Entity Resolution POC Report – Supplier Database Cleansing for Procurement Digitalization

By Vlădeanu Ionuț

Table of Contents

Introduction

This report presents the results of a Proof of Concept (POC) simulation addressing a common challenge faced by procurement teams: supplier databases that are cluttered with duplicates, outdated records, and inconsistent entries. Such issues hinder spend analysis, complicate supplier negotiations, and slow down strategic decision-making.

The client, a large manufacturing company, is currently piloting solutions from two providers as part of its digitalization journey. The objective of this exercise was to demonstrate how entity resolution can improve supplier data quality by reconciling messy input records with real-world companies.

Our task was fourfold:

1. **Entity Resolution** – For each of the 591 input suppliers, select the single best real-world entity among up to five candidate matches provided by Veridion. If no suitable candidate exists, the row should be marked as **UNMATCHED**.
2. **Data Analysis, Cleaning, and Quality Control** – Beyond the core matching task, we also performed basic data cleaning (normalizing company names, removing legal suffixes, standardizing URLs) to reduce noise and inconsistencies. In parallel, we reviewed the consistency of returned attributes (names, countries, websites, industries, etc.) and flagged gaps or anomalies that could affect downstream procurement workflows.
3. **Summarization** – Document the methodology, decision criteria, and outcomes of the project in a transparent and auditable way, highlighting both strengths and limitations of the approach.
4. **Publication** – Package the work in a structured format suitable for client review, demonstrating both the logic applied and the potential business impact.

The analysis below follows this structure, detailing the logic applied, the engineered features, examples of successful and failed matches, and the overall distribution of results. While the algorithm achieves a high degree of alignment with human judgment, known edge cases remain, reinforcing the importance of human-in-the-loop validation for low-confidence outcomes.

Entity Matching Report – Methodology and Results

Objective

For each client input row (one supplier), the task is to select a single best real-world entity among up to five Veridion candidates. If none are valid, the row is marked **UNMATCHED**. Since manually reviewing all 591 inputs (2,951 candidate rows) is impractical, I designed a transparent scoring algorithm that replicates the reasoning of a human analyst, supported by spot-checks.

Human Matching Criteria

The matching process follows a structured set of rules that reflect how an analyst would decide:

1. **Name (mandatory)**
   Compare `input_company_name` with `company_name`, `company_legal_names`, and `company_commercial_names`. Accept exact matches, legal vs. commercial variants, and common abbreviations (after normalization).
2. **Country (near-mandatory)**
   `input_main_country/_code` should match the candidate's `main_country/_code`.
3. **City/Region (supporting signal)**
   Prefer candidates whose `main_city/main_region` or listed locations correspond to the input locality.
4. **Website/Domain (very strong signal)**
   A matching official domain (e.g., `apple.com` vs. `apple.com/ro`) strongly validates the brand.
5. **Industry Coherence (NAICS/NACE/SIC/Sector)**
   The company's industry must align logically with its name (e.g., telecom vs. real estate).
6. **Size Sanity Check**
   Revenue and employee count are used only to identify unrealistic candidates (implausibly large or small compared to context).

If the name has no relationship, the result is **UNMATCHED** even if country or city are correct. If the country differs clearly, the result is also **UNMATCHED**, unless the input describes a global brand and the candidate represents its local legal entity.

Algorithmic Implementation

Step 1 — Helper Functions

- **normalize_name**: convert to lowercase, replace `&` with `and`, strip punctuation, collapse spaces, and remove legal suffixes (Inc, Ltd, GmbH, Sdn Bhd, etc.).
- **name_similarity(a, b)**: compute token-sorted similarity (independent of word order).
- **country_match(row)**: check ISO code or country string.
- **city_region_match(row)**: verify whether the input city/region matches the candidate's `main_city/main_region` or appears in locations.
- **website_match(row)**: normalize URLs (remove http/https/www) and test for containment.

These proxies emulate human judgment: prioritize name, then country, then website, with locality as an additional supporting factor.

Step 2 — Candidate-Level Features

For each candidate row:

- `name_score` = best similarity vs. company_name, legal_names, or commercial_names
- `country_match`, `city_region_match`, `website_match` = boolean values derived from helper functions

Step 3 — Weighted Total Score

Each feature contributes to a weighted score, reflecting practical importance:

- Name = 0.60 (primary identifier)
- Country = 0.25 (nearly mandatory)
- Website = 0.20 (brand confirmation)
- City/Region = 0.10 (tie-breaker)

`total_score = 0.60*name + 0.25*country + 0.20*website + 0.10*city` (clipped at 1.0).

Step 4 — Candidate Classification

A candidate is labeled **CANDIDATE_MATCH** if any of the following conditions hold:

- `name_score ≥ 0.80` and `country_match`
- `website_match` and `country_match` and `name_score ≥ 0.50`
- `total_score ≥ 0.85`
- `0.65 ≤ name_score < 0.80` with `country_match` and (website_match or city_region_match)

Otherwise, the candidate is labeled **CANDIDATE_UNLIKELY**.

Step 5 — Best Candidate Selection

For each `input_row_key`:

- Prefer any **CANDIDATE_MATCH**; otherwise, select the candidate with the highest `total_score`.
- Generate an interpretable `match_reason`, e.g.:
  `name=exact, country ok, city ok, website ok`

Final output includes:
`best_veridion_id, best_company_name, best_total_score, best_name_score, best_candidate_status, match_reason.`

Step 6 — Final Input-Level Status

- **MATCHED** if `CANDIDATE_MATCH` and `total_score ≥ 0.70`
- **UNMATCHED** if `total_score < 0.55`
- **BEST_GUESS** otherwise

This produces a clear "green/amber/red" decision framework for category managers.

Engineered Columns

To ensure full auditability and filtering:

- `name_score`, `country_match`, `city_region_match`, `website_match`
- `total_score`, `candidate_status`, `match_reason`, `match_status`

Threshold Justification

- **0.80 name_score**: equivalent to exact brand/legal alias.
- **0.65–0.80 name_score**: "close variant," acceptable only with supporting evidence.
- **0.55 total_score**: below this, false positives increase significantly → UNMATCHED.
- **0.70 total_score**: aligned well with manual decisions → chosen as MATCHED threshold.

Examples: Successes and Failures

Successful Cases

- Exact or legal/commercial variants with same country → high score, reliable match.
- Slight name variants corroborated by website → correctly upgraded to MATCH.

Failure Modes (Manual QC Findings)

- **Block 0 — 24-SEVEN MEDIA NETWORK (PRIVATE) LIMITED**
  None of the five candidates were correct. Algorithm still selected *New Millennium Network* as "least wrong."
  Correct: **UNMATCHED**.
- **Block 32 — AMDOCS DEVELOPMENT LIMITED (Cyprus, Limassol)**
  All five candidates were generic *Amdocs* entities (US/Canada). One had the correct global site `amdocs.com`, but the country mismatch forced rejection.
  Correct: **MATCHED** (brand + website), despite missing Cyprus in Veridion.
- **Block 106 — CONSCIA DANMARK A/S**
  Matched by name to "Conscia Danmark A/S," but website pointed to `laugusen.io` (unrelated). Correct entity was *Conscia Moments* with `conscia.com`.
- **Blocks 107–108 — CONTROL RISKS GROUP LIMITED**
  Matched to "Control risks" with `risks-group.com` (wrong). Correct entity: *Control Risks Group Holdings Ltd.* with `controlrisks.com`.

Takeaway: when all five candidates are incorrect, a relative scorer still selects one. Maintaining **UNMATCHED** as a valid outcome is essential, with human review recommended for low-confidence cases.

Results Summary

Out of 591 input companies, the algorithm produced the following distribution:

- **89.2% MATCHED** (527 inputs)
- **9.8% BEST_GUESS** (58 inputs, good candidates but weaker corroboration)
- **1.0% UNMATCHED** (6 inputs, irrelevant or insufficient candidates)

These percentages align with the overall trends observed, but they are not fully error-proof. As highlighted in the failure cases, certain matches were misclassified due to incomplete or misleading candidate data. Therefore, results should be interpreted as **approximate indicators rather than absolute ground truth**, and human review remains necessary for low-confidence cases.

Recommendations for Production Hardening

- **Website trust weighting**: down-weight domains that are newly registered or unrelated (heuristics based on TLD or suspicious keywords).
- **Name blocking**: allow explicit "NOT THIS" lists per input to prevent false relative matches.
- **Geo disambiguation**: prioritize candidates whose location data includes both input city and region.
- **Active learning**: feed corrections from manual reviewers into threshold calibration and suffix lists.

This report documents both the methodology and practical results of the entity-matching algorithm. The framework balances automation with interpretability, while leaving room for human oversight where algorithmic confidence is low.

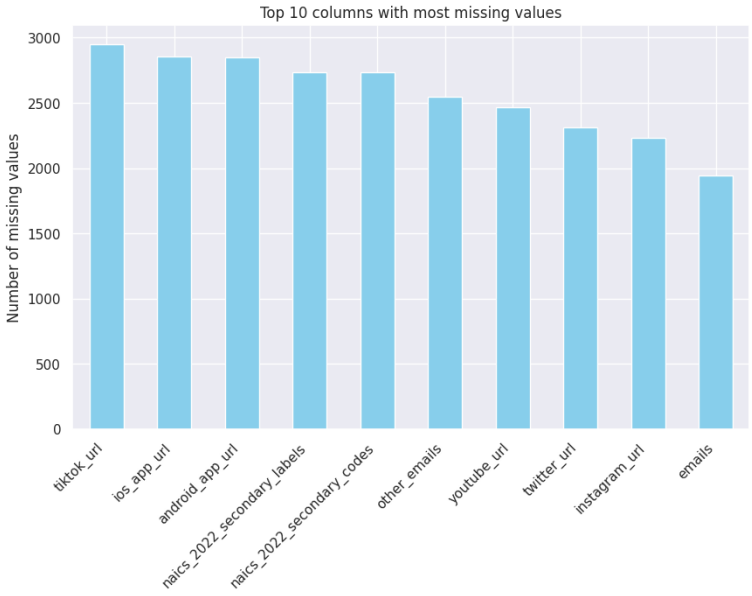
Data Analysis, Cleaning, and Quality Control

In parallel with entity resolution, I conducted a systematic review of the dataset to evaluate its completeness, consistency, and reliability. The goal was not to correct every issue, but to diagnose potential risks in data quality and ensure that outputs are interpretable and actionable for procurement.

1. Missing Values

A column-level assessment revealed significant gaps across multiple attributes. While key identifiers such as `input_row_key`, `input_company_name`, and `veridion_id` were complete, a large share of descriptive and classification fields contained missing entries:

- **Company attributes**: `company_legal_names` (1,380 missing), `year_founded` (1,708 missing).
- **Financial attributes**: `revenue` (1,385 missing), `employee_count` (1,291 missing).
- **Industry codes**: Secondary NAICS/NACE/ISIC labels are missing in more than 2,700 cases.
- **Contact information**: `emails` (1,946 missing), `other_emails` (2,546 missing), `website_url` (671 missing).
- **Social media links**: `tiktok_url` (2,951 missing), `ios_app_url` (2,855 missing), `android_app_url` (2,848 missing).

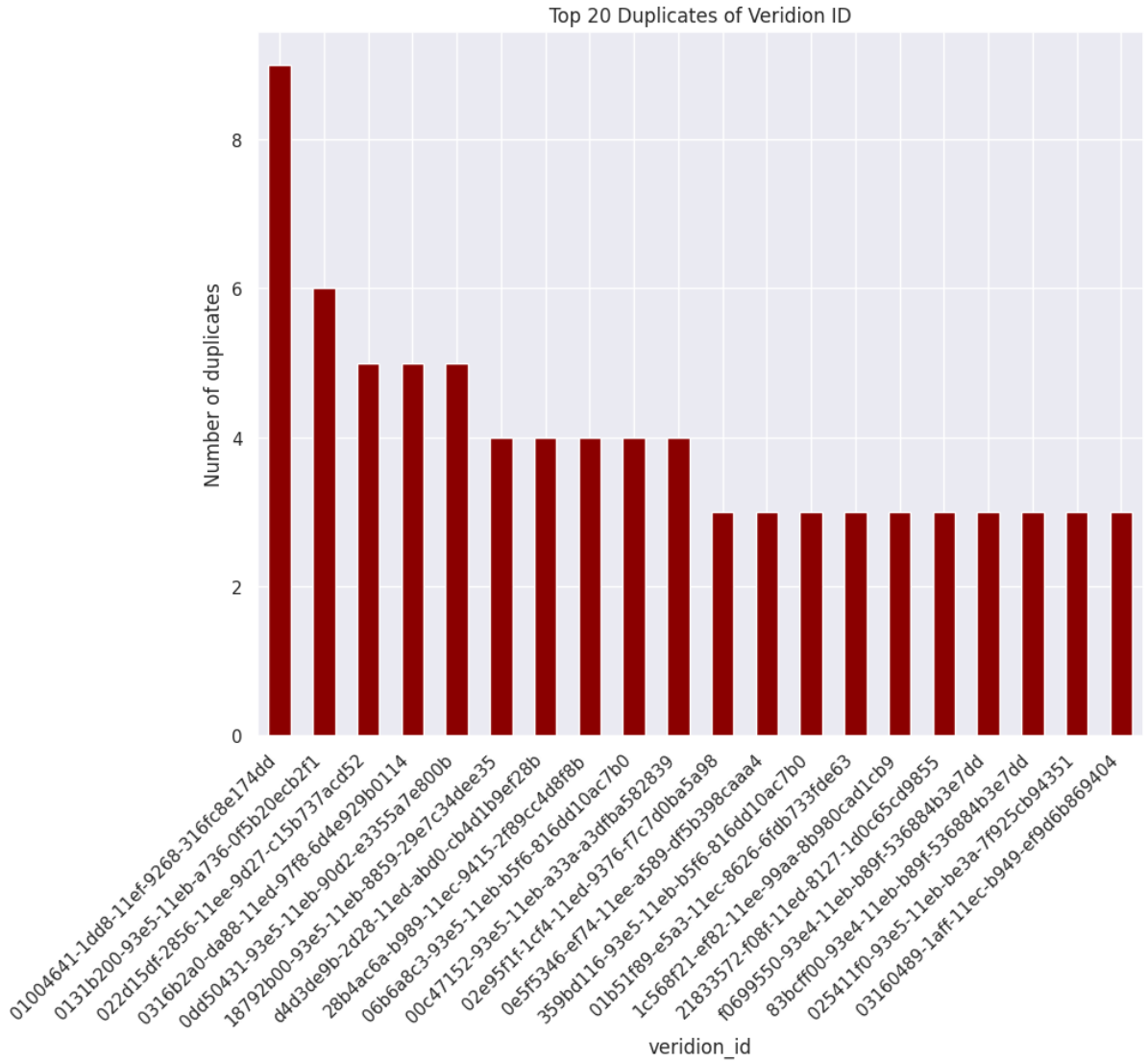


The plot "Top 10 columns with most missing values" confirms that mobile app URLs and social media handles are the least populated fields, while financial and classification attributes also show substantial sparsity. These gaps reduce the dataset's utility for advanced analytics, particularly in supplier profiling and segmentation.

2. Duplicates

Duplicate checks on `veridion_id` revealed **180 duplicated IDs**, some repeated up to 9 times. Notable examples include IDs linked to *Huawei*, *Telenor*, and *Zalaris*.

- **Interpretation**: These duplications often represent multiple subsidiaries, alternate names, or regional entries tied to the same global entity.
- **Risk**: Without deduplication, spend analysis or supplier counts would be inflated, potentially misleading procurement stakeholders.



The bar plot “Top 20 Duplicates of Veridion ID” illustrates the skew, with several entities appearing four to six times.

3. Outliers

To validate numeric consistency, I examined the distribution of `revenue` and `employee_count`:

- **Revenue**: Strongly right-skewed, with outliers exceeding USD 10 trillion.
- **Employees**: Most firms have fewer than 10,000 employees, but a few report over 1.5 million.
- **Notable outliers**: *Amazon*, *Deloitte*, *ISS*, and *Huawei*.

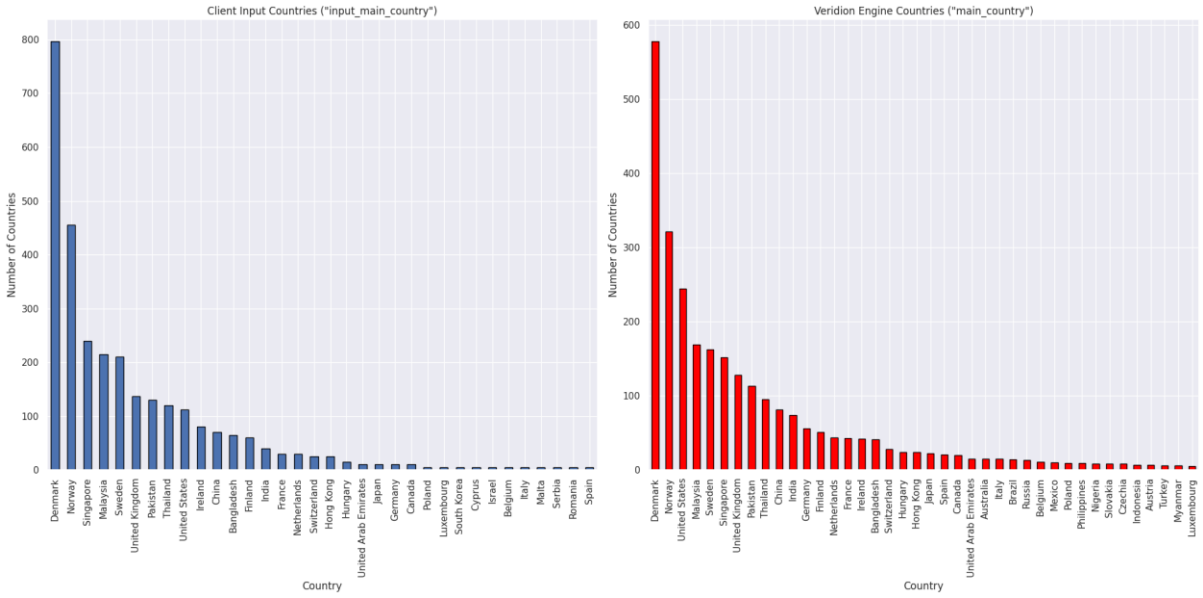
Scatter plots (log–log scale) confirm the expected positive correlation between workforce size and revenue, but extreme anomalies suggest possible estimation or scraping errors.

- **Conclusion**: Numeric fields are informative but require QC before being used for financial benchmarking or supplier tiering.

4. Distribution Checks

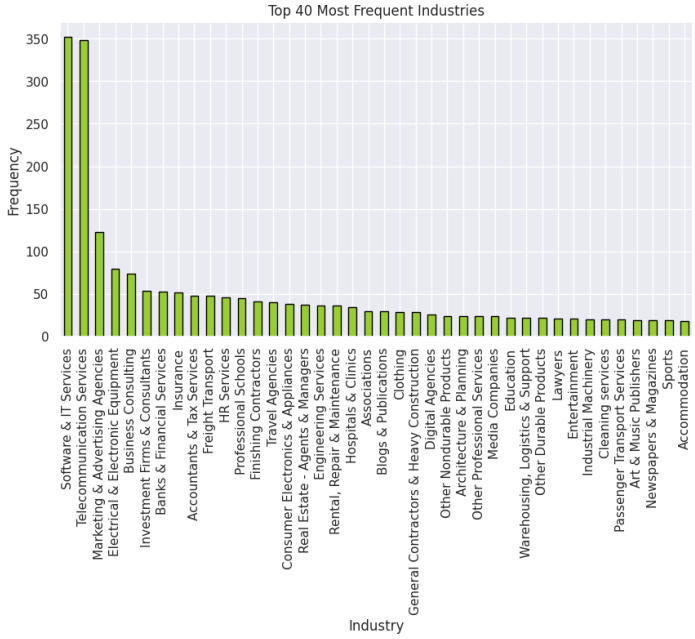
Countries

- Input records are concentrated in Denmark, Norway, and Singapore.
- Veridion mappings skew more heavily towards the United States, indicating normalization differences.
- Some long-tail countries (e.g., Romania, Spain, Serbia) are underrepresented.



Industries

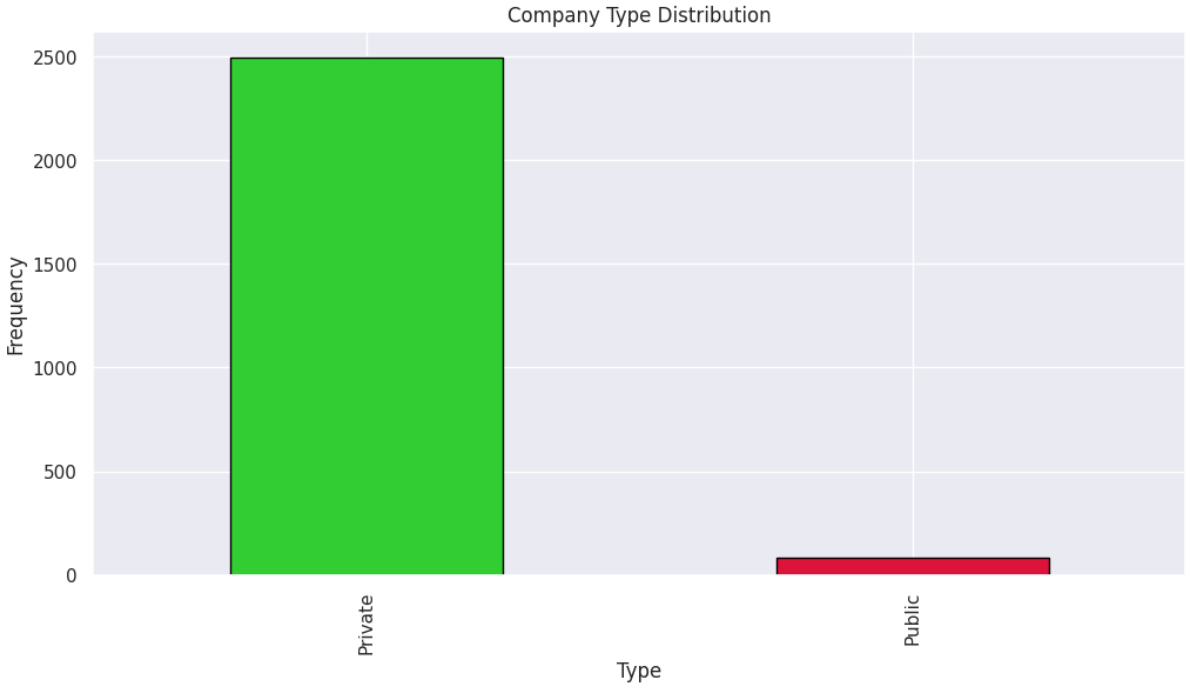
- Most frequent categories: IT Services, Telecom, Marketing & Advertising.
- Strong concentration in digital sectors, while manufacturing and logistics are less represented.



These distribution checks highlight **coverage biases**: the dataset is strongest in IT/telecom industries and Nordic/Asian geographies, while being less complete in other regions and verticals.

Company Type

- Private companies dominate (≈95%), with public companies making up a very small share.



5. Cross-Validation Logic (Attribute Consistency)

Consistency checks were performed across key attributes:

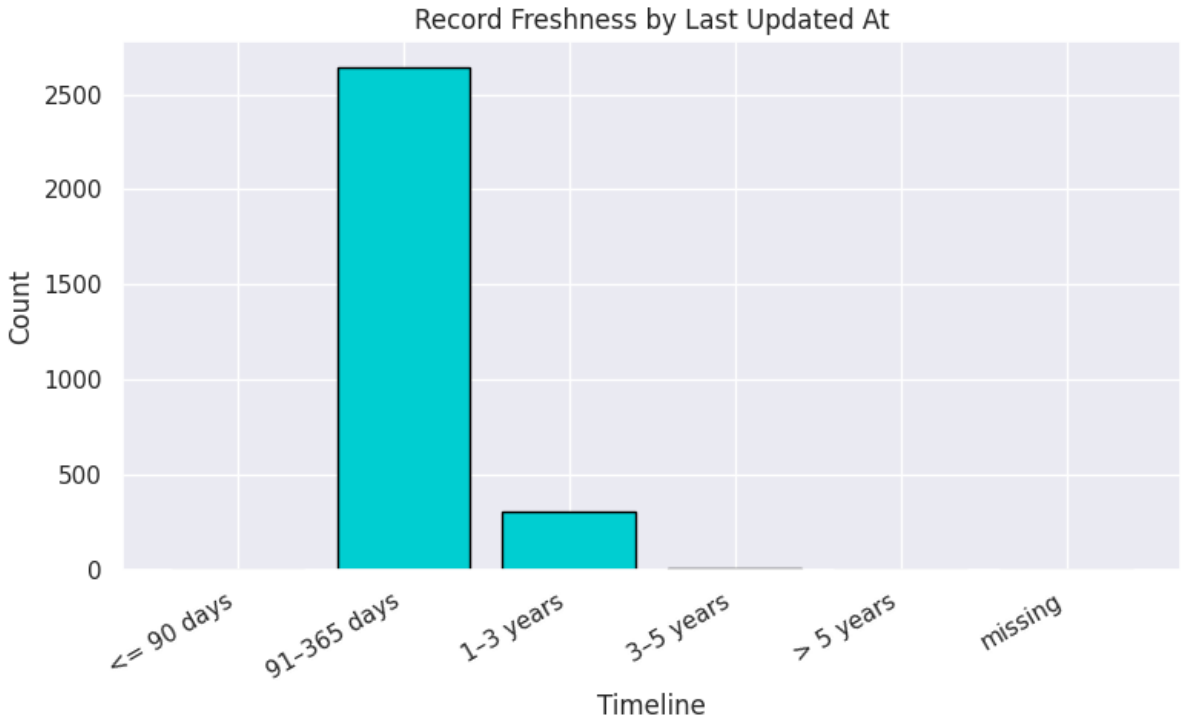
- **Employee Count**: 832 records use modelled estimates.
- **Revenue**: 1,184 records use modelled values, which should be treated as approximations.
- **Website Data**: No inconsistencies found between `website_url`, `website_domain`, and `website_tld`, suggesting high reliability.

Overall, while financial and HR-related fields require cautious interpretation, website and domain-level information is robust and trustworthy.

6. Timeliness Check

Record freshness was assessed using `last_updated_at`:

- ~90% updated in the past year.
- ~10% updated 1–3 years ago.
- None older than 5 years.
- 0% updated in the last 90 days.



Interpretation: The dataset is relatively recent but not fresh. For procurement-critical suppliers, updates within ≤90 days would be recommended.

7. Coverage Metrics – Supplier Contact Information

Coverage of supplier communication channels is uneven:

- **Website**: ~77% available
- **Primary Email**: ~61% available
- **Phone Numbers**: ~75% available
- **LinkedIn**: ~45% available
- **Social Media (Twitter, Instagram, TikTok)**: <25% available

This suggests that while most suppliers can be contacted via official websites and phone numbers, digital engagement data (social media, apps) is sparse. For procurement use cases, the dataset is therefore better suited to **formal supplier engagement** rather than digital marketing analysis.

8. Data Cleaning Summary

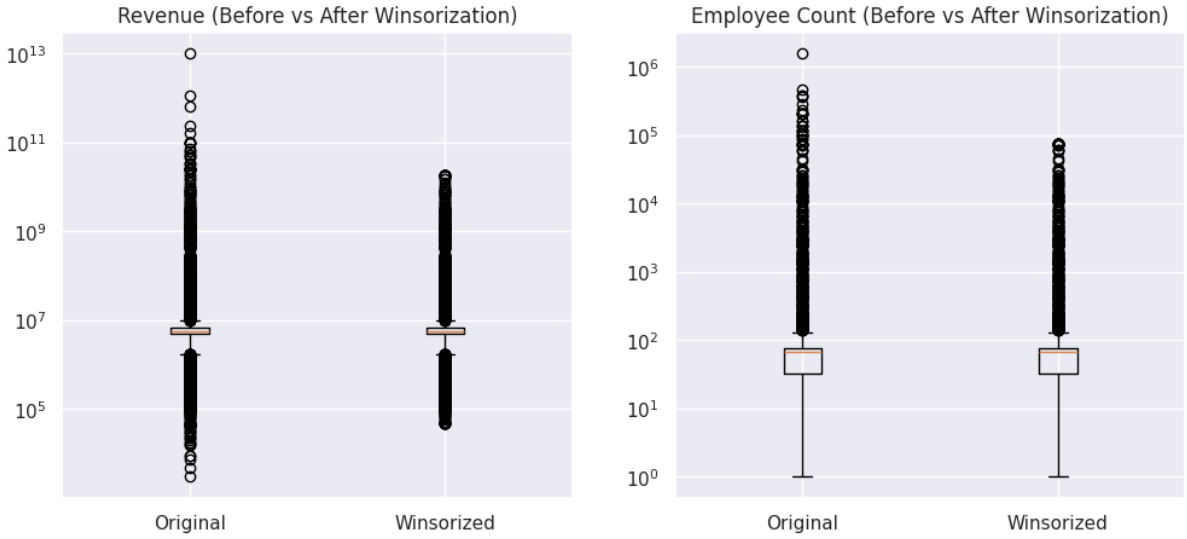
In addition to analysis and QC, several targeted data cleaning steps were applied to improve dataset quality and ensure consistency. The goal was not to fully productionize the dataset, but to address the most impactful issues while preserving transparency for downstream use.

1. Missing Values Imputation

- Imputed missing values for numeric fields (`revenue`, `employee_count`, `year_founded`) using the median.
- Imputed missing values for categorical fields (e.g., `company_type`) using the mode.
- Created flag columns (e.g., `revenue_imputed`) to clearly mark imputed records.
  This approach maintains usability of the dataset while ensuring that imputed values remain distinguishable from reported ones.

2. Outlier Treatment

- Applied winsorization at the 1st and 99th percentiles for `revenue` and `employee_count`.
- Outliers were capped rather than removed, minimizing the influence of extreme values without discarding potentially valid records.
- This adjustment preserved the overall distribution shape while reducing distortions in statistical analysis.



3. Deduplication

- Identified duplicate companies based on `veridion_id`.
- Retained the first occurrence and removed subsequent duplicates, ensuring that each company is represented only once.
- This prevents bias in aggregation tasks (e.g., revenue sums, supplier counts).

Next Steps (Not Implemented Here)
For a production pipeline, additional steps would include:

- Standardizing categorical values (harmonizing industry labels).
- Parsing and normalizing address fields (street, postcode, region).
- Validating external links (websites, social media) with regex or APIs.
- Handling multi-location companies by expanding and normalizing the `locations` attribute.

For the scope of this POC, the three steps above provided sufficient improvements to support analysis and ensure transparency.

Overall Summary

The data analysis and quality control exercise identified several strengths and weaknesses in the supplier dataset:

- **Strengths**:
  - Core identifiers (company name, country, website) are highly reliable.
  - Contact coverage is strong for websites and phone numbers.
  - Website fields are internally consistent and trustworthy.
  - Most records have been updated within the past year.
- **Weaknesses**:
  - Missing values remain a challenge, particularly for financial and HR data.
  - Duplicates and outliers introduce noise into aggregated results.
  - Industry and geography distributions are biased, with concentration in IT/telecom sectors and Nordic/Asian countries.
  - Social media and app coverage is very sparse, limiting digital engagement use cases.
  - Revenue and employee counts often rely on modelled values rather than official data.

Overall, the dataset is **fit for purpose** in procurement entity resolution and spend consolidation, but it requires careful handling of outliers, duplicates, and missing fields. Human-in-the-loop validation remains important for edge cases, especially where company identity or financials are critical.

Key Recommendations

Based on the entity resolution process and the data analysis, cleaning, and quality control review, the following recommendations are proposed for improving supplier data quality and ensuring readiness for procurement use cases:

1. **Strengthen Data Freshness**
   - Establish automated update cycles, ideally ≤90 days for strategic suppliers.
   - Prioritize refreshing financial and HR-related attributes (revenue, employee count), as these degrade fastest in relevance.
2. **Enhance Data Completeness**
   - Enrich missing values for critical fields (year_founded, industry codes, revenue, employees) through external data sources or targeted enrichment pipelines.
   - Expand coverage of contact attributes, especially LinkedIn and primary emails, to enable more effective supplier engagement.
3. **Refine Deduplication and Identity Resolution**
   - Implement stricter deduplication rules to prevent inflation of supplier counts.
   - Consolidate subsidiaries and alternate names under parent entities where relevant, to provide category managers with a unified supplier view.

4. **Improve Outlier Handling**
    - Incorporate automated detection and treatment of extreme revenue and employee values into the pipeline.
    - Clearly flag values that are modelled or estimated, distinguishing them from reported figures.
5. **Balance Industry and Geographic Coverage**
    - Proactively source additional suppliers in underrepresented industries (e.g., manufacturing, logistics) and regions (outside Nordic and Asian concentrations).
    - This will improve representativeness and reduce bias in spend analyses.
6. **Human-in-the-Loop Validation for Low-Confidence Cases**
    - Maintain manual review checkpoints for edge cases flagged as **BEST_GUESS** or **UNMATCHED**.
    - Feedback from reviewers should feed into active learning mechanisms to continuously improve thresholds and rules.

With these actions, the dataset can move from a functional proof-of-concept toward a **production-ready supplier intelligence asset**, supporting not only spend consolidation and cost-saving strategies, but also longer-term goals such as supply chain sustainability and resilience.