Data Insights - Key Takeaways.

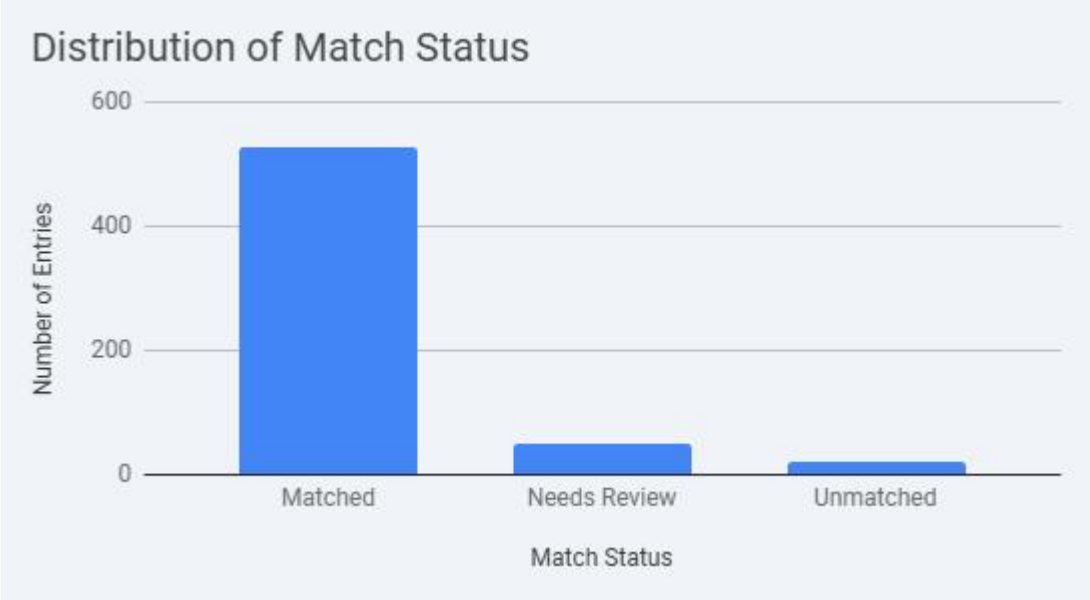
Geographical Concentration

A significant portion of the companies in the dataset are concentrated in a few key countries, with Denmark leading the way.
Denmark accounts for the largest share of companies, representing 19.59% of the total.
Norway and the United States follow with 10.88% and 8.27% respectively, indicating a strong presence in these regions as well.
Other notable countries include Malaysia, Sweden, Singapore, and the United Kingdom, each contributing a substantial percentage to the overall company count.

Dominant Business Categories

The technology and telecommunications sectors emerge as prominent business categories within the dataset.
"Tech Companies" and "Telecommunications" are the most frequently occurring business categories, each with over 180 companies.
"Software Development Services" also shows a strong presence, indicating a robust market for software-related businesses.
Other significant categories include "Cable & Internet Providers," "Hosting & Cloud Infrastructure Providers," and "Business Consulting."

Key Finding - High Match Rate and Quality



The dataset contains 592 entries and 21 columns, providing information about companies, their locations, matched company details, and match scores.

Here are some additional interesting findings from the data:

Match Score Distribution for 'Needs Review' and 'Unmatched' Entries

To understand the characteristics of entries that are not fully matched, I'll examine the match_score for Needs Review and Unmatched entries.

The match_score for 'Needs Review' entries ranges from 75 to 84.7, with an average of 80.71. This suggests that even entries requiring review often have relatively high match scores, indicating that the review might be needed for subtle discrepancies rather than outright mismatches. For 'Unmatched' entries, the match_score ranges from 41.4 to 73.3, with an average of 60.42. This lower average score is expected, as these entries did not find a definitive match.

Discrepancies in Country Information for 'Needs Review' and 'Unmatched' Entries

For 'Needs Review' entries, 11 out of 47 (23.40%) have a mismatch between input_main_country and main_country. For 'Unmatched' entries, 11 out of 20 (55.00%) have a mismatch between input_main_country and main_country. This indicates that country discrepancies are a significant factor contributing to 'Unmatched' status and a notable factor for 'Needs Review' status.

Unique Matched Companies

There are 585 unique matched_veridion_id entries in the dataset. This means that while there are 592 total entries, some input companies might have been matched to the same Veridion ID, or some entries were not matched at all.

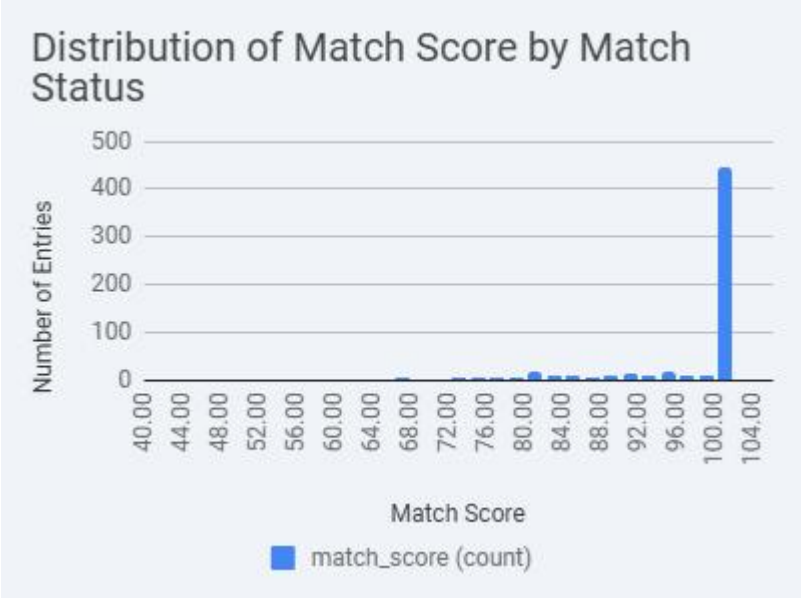
Distribution of Match Score by Match Status

To provide a comprehensive overview, I'll analyze the distribution of match_score across all match_status categories: 'Matched', 'Needs Review', and 'Unmatched'.

The match_score distribution varies significantly across match statuses:

> **Matched**: The scores are very high, with a minimum of 85 and an average of 98.95, indicating strong matches.
> **Needs Review**: The scores are also relatively high, ranging from 75 to 84.7, with an average of 80.71. This suggests that these entries are close to being matched but require human verification.
> **Unmatched**: The scores are considerably lower, ranging from 41.4 to 73.3, with an average of 60.42, which is expected for unmatched entries.

These findings provide a deeper understanding of the matching process and the characteristics of different match statuses.



Investigation of 'Company Legal Names'

For 'Matched' entries: 196 out of 525 (37.33%) have null company_legal_names.
For 'Needs Review' entries: 24 out of 47 (51.06%) have null company_legal_names.
For 'Unmatched' entries: 8 out of 20 (40.00%) have null company_legal_names.

This is an interesting finding:

'Needs Review' entries have the highest percentage of missing legal names, which could be a contributing factor to why they require manual review.
'Unmatched' entries also have a relatively high percentage of missing legal names.
Even 'Matched' entries have a portion of missing legal names, suggesting that the presence of legal names is not a strict requirement for a successful match, but its absence might increase the complexity or uncertainty of the match.

Relationship between 'Review Needed' and Match Score

Entries that do not require review (review_needed = False) have a significantly higher average match_score of 97.53, compared to entries that do require review (review_needed = True), which have an average match_score of 80.71. This suggests

that the review_needed flag effectively identifies matches with lower confidence for human verification.

Investigation of 'Company Legal Names

The analysis of company_legal_names reveals that missing legal names are more prevalent in 'Needs Review' entries (51.06%) and 'Unmatched' entries (40.00%) compared to 'Matched' entries (37.33%). This suggests that the absence of legal names might contribute to the difficulty in achieving a definitive match or a confident match.

Completeness of Contact Information

LinkedIn URL:
'Matched': 284 out of 525 (54.10%) have null LinkedIn URLs.
'Needs Review': 28 out of 47 (59.57%) have null LinkedIn URLs.
'Unmatched': 10 out of 20 (50.00%) have null LinkedIn URLs.

Website URL:
'Matched': 137 out of 525 (26.10%) have null Website URLs.
'Needs Review': 10 out of 47 (21.28%) have null Website URLs.
'Unmatched': 2 out of 20 (10.00%) have null Website URLs.

The completeness of linkedin_url and website_url does not show a clear pattern where matched companies consistently have more complete contact information. For linkedin_url, 'Needs Review' entries have the highest percentage of missing URLs (59.57%), followed by 'Matched' (54.10%) and then 'Unmatched' (50.00%). For website_url, 'Unmatched' entries have the lowest percentage of missing URLs (10.00%), while 'Needs Review' has 21.28% and 'Matched' has 26.10%. This indicates that the presence of these URLs alone might not be a primary factor for determining match status.

Geographic Distribution of 'Needs Review' and 'Unmatched' Entries

A significant number of 'Needs Review' and 'Unmatched' entries originate from Norway (17 entries) and Denmark (16 entries), which are also among the top countries for matched companies. This suggests that while these countries have many companies in the dataset, they also present challenges for automated matching.

Specific regions and cities such as 'Capital Region Of Denmark' (10 entries), 'Oslo' (8 entries), 'Singapore' (7 entries), and 'Copenhagen' (7 entries) appear frequently in the problematic entries. This highlights specific geographic areas where the matching algorithm might encounter more difficulties, warranting further investigation into the data quality or unique attributes of companies from these locations.

Summary of Findings from Duplicate `veridion_id` Analysis

The analysis reveals several instances where different input companies are matched to the same `matched_veridion_id`. This can occur due to various reasons, as observed in the examples:

- Variations in Company Name:
  `3 STEP IT A/S` and `3 STEP IT AS` both matched to `3 Step IT A/S.`.

This indicates that minor variations in company names (e.g., presence/absence of legal suffixes like A/S, AS) can lead to different input entries mapping to the same actual company.

- Related Entities/Subsidiaries:
  `TELENOR GO PTE. LTD.` and `TELENOR PROCUREMENT COMPANY PTE. LTD.` are both matched to `Telenor Inpli Singapore`. This suggests that different entities or subsidiaries of a larger group might be consolidated under a single `veridion_id` if they are considered part of the same core business or legal entity by the matching system.

  `TELENOR ASA` and `TELENOR REAL ESTATE AS` both matched to `Telenor`. This is another example where different operational units or related companies under a common brand are linked to the same parent `veridion_id`.

`DELOITTE ADVOKATFIRMA AS` and `DELOITTE AS` both matched to `Deloitte Norge`. This highlights matching to a parent entity even when the input specifies a specific branch or legal form.

- Regional/Jurisdictional Variations of the Same Company:
  `SQUARETRADE EUROPE LIMITED` and `SQUARETRADE LIMITED` both matched to `SquareTrade`. This indicates that different legal entities operating in different regions or with slightly different legal forms, but representing the same core company, are mapped to a single `veridion_id`.
  `GOOGLE COMMERCE LIMITED` and `GOOGLE IRELAND LIMITED` both matched to `Google Commerce Limited`. This is another instance of regional variations of a global company being mapped to a single ID.

In some cases, one of the duplicate matches might have a lower `match_score` or be flagged as `Needs Review` or even `Unmatched`, while another input for the same `veridion_id` is `Matched` with a high score. For example, `SoftwareONE Co., Ltd.` (Needs Review, score 75.9) and `SOFTWAREONE PTE. LTD.` (Unmatched, score 73.3) are both attempting to match to `SoftwareOne`. This suggests that while a `veridion_id` exists, the quality of the input data or slight variations can lead to different matching outcomes for what might be the same underlying entity.

The occurrence of duplicate `veridion_id` for different input companies is often a result of the matching algorithm's attempt to consolidate various representations of the same real-world entity under a single unique identifier. This is a common challenge in data matching and entity resolution, driven by: Data Inconsistencies. Variations in company names (e.g., abbreviations, legal suffixes, typos), different registered legal entities for the same operational business, or different branches/subsidiaries of a parent company.

Matching Logic: The algorithm might be designed to be broad enough to capture related entities, assuming they refer to the same core business for the purpose of the `veridion_id`.

Input Data Granularity: The input data might contain entries for various legal forms or operational units of a single larger organization.

Solutions:

Addressing duplicate `veridion_id` depends on the desired outcome and the specific use case of the data. Here are some potential solutions:

1. **Refine Matching Rules (Pre-processing):**
   Implement pre-processing steps to normalize company names (e.g., remove common legal suffixes, standardize abbreviations, correct common typos) before feeding them to the matching algorithm.
   If the goal is to distinguish between parent companies and their subsidiaries, a more sophisticated hierarchical matching approach might be needed, where the `veridion_id` represents the ultimate parent, and additional fields indicate the specific subsidiary.

2. **Post-Processing and Deduplication:**

   Prioritization of manual review for duplicate `veridion_id` where one or more entries are 'Needs Review' or 'Unmatched'. This allows human judgment to determine if they truly refer to the same entity or if the match is incorrect.
   Consolidation Strategy: Define a strategy for consolidating duplicate `veridion_id` entries. For example, if multiple input companies map to the same `veridion_id` with high match scores, they can be treated as valid matches to the same entity. If the goal is to have a unique `veridion_id` for each input company, then the matching criteria might need to be stricter.
   Leverage Additional Data Points: Use other fields like `main_country`, `main_region`, `main_city`, `website_url`, or `linkedin_url` in conjunction with `company_name` to refine matching and differentiate between similar entities.

3. **Feedback Loop to Matching Algorithm:**
   We can use the insights gained from manual reviews of duplicates to refine and improve the matching algorithm, making it more intelligent in distinguishing between truly distinct entities and different representations of the same entity.