ENTITY RESOLUTION

ASSESSMENT FOR VERIDION

OLARU MARIA MADALINA

WORKFLOW

Filtering Parameters

Data Normalization

Grouping Entities

WHAT TRULY DEFINES A COMPANY?

The answer depends primarily on the type of analysis we want to perform—whether we are looking at a corporate group or a legal entity. Based on the data received, I will focus on analyzing **Legal Entities**.

→ Unique registration code

→ Company name

→ Country

→ Phone number

→ Email

→ Website

→ Adress

→ Main activity

Checking for null values in each column

Remove columns with over 90% missing values

Choosing grouping parameters

Remove similar columns

Eliminate non-essential or technical metadata

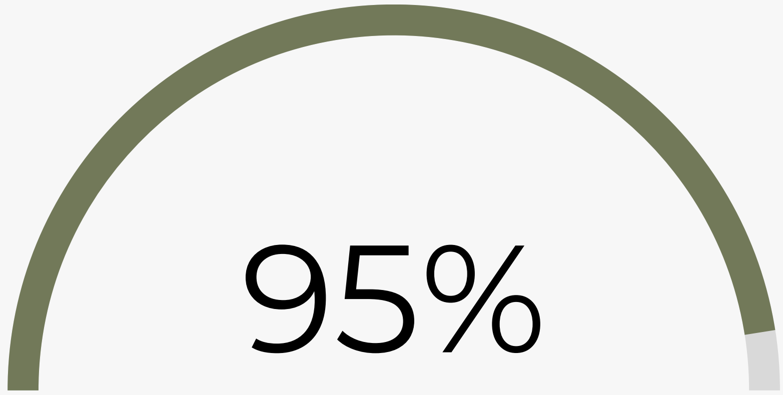
WHAT ARE THE KEY ATTRIBUTES IN THE DATASET THAT WE CAN USE?

Even if a certain attribute is very effective in distinguishing between two records, if it is not sufficiently populated, it cannot be relied upon as a primary feature. However, it can be very useful when the initial grouping does not yield conclusive results.

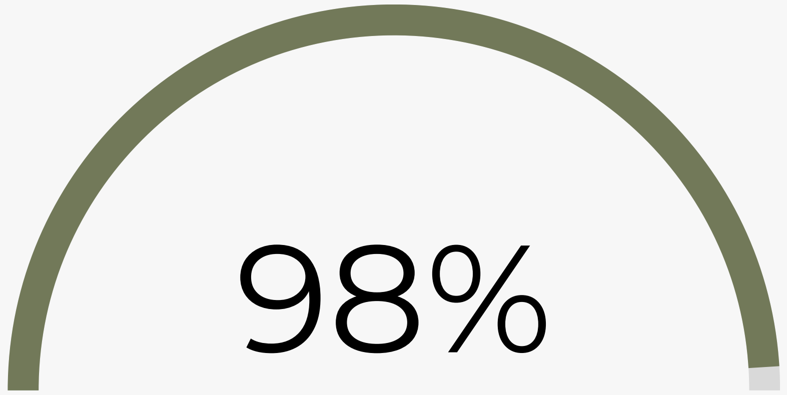
Remove similar columns - main examples

- company_name vs company_commercial_names vs company_legal_names → Keep company_name because it has the most populated values and the others may be misleading
- main_country_code vs main_country → Keep main_country_code as it's easier to filter and the values are standardized
- website_url, website_domain, website_tld → Keep website_domain is partially cleaned and contains the most important information

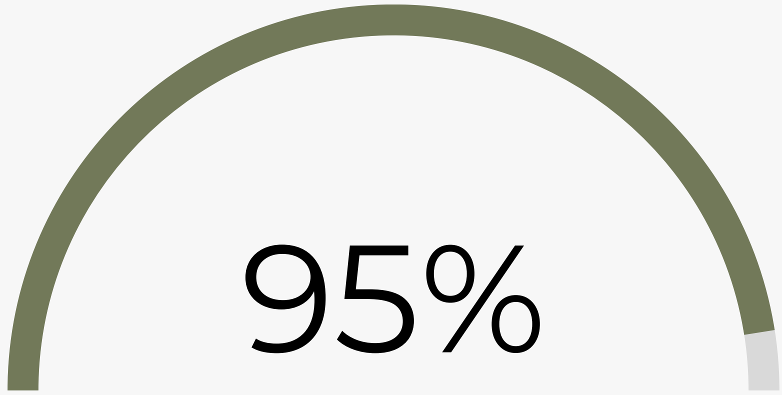
ESSENTIAL PARAMETERS



main_country_code



company_name



website_domain

FILTERING PARAMETERS

THOUGHT PROCESS BEHIND MY SELECTION

main_country_code →

High impact on entity uniqueness - We may have two entities with the same name but in different countries.
Strong filtering dimension
Data completeness and derivability - Even if some values are missing, country information can often be inferred from other fields

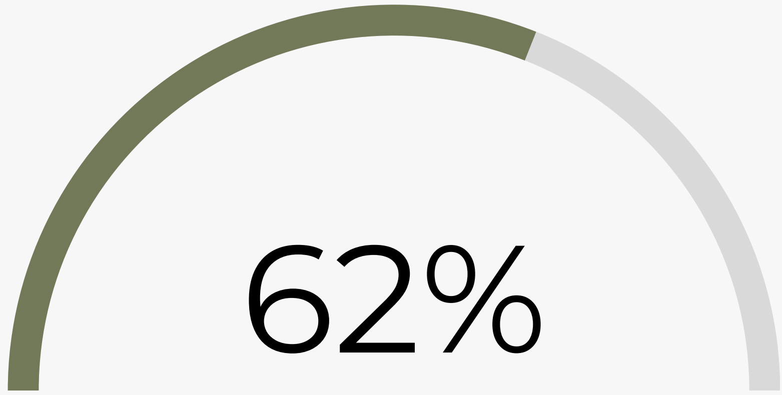
company_name →

High availability across dataset
Strong signal for matching
Primary identifier

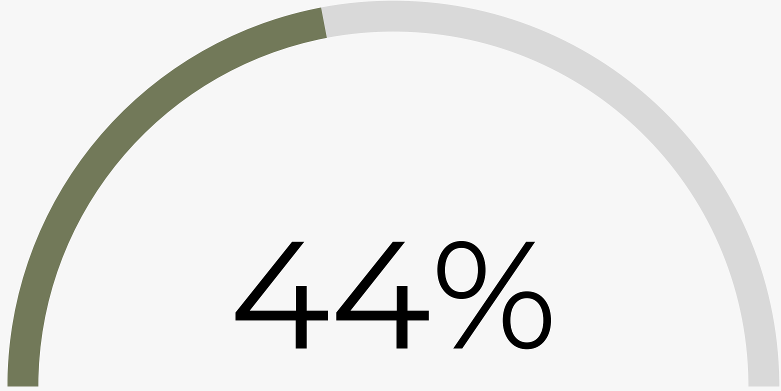
website_domain →

Are typically unique to each company
Difficult to duplicate across entities
Often linked to official identity

SECONDARY PARAMETERS



primary_phone



facebook_url

THOUGHT PROCESS BEHIND MY SELECTION

primary_phone →

Going beyond the main parameters, the telephone number, although to a lesser extent, can be defining but also very specific in terms of grouping entities.

facebook_url →

If we have made sure they are in the same country and the company name is similar, the facebook URL can be a good indicator to identify if we are talking about the same entity.

WHY CERTAIN ATTRIBUTES WEREN'T USED?

Main business activity

The data could be outdated by up to two years (based on last_updated_at) and companies may have changed sectors but also the domain field can vary slightly depending on the data source.

Address

When a company operates multiple branches within the same country, filtering by address may inadvertently split a single entity into two or more distinct records.

DATA NORMALIZATION

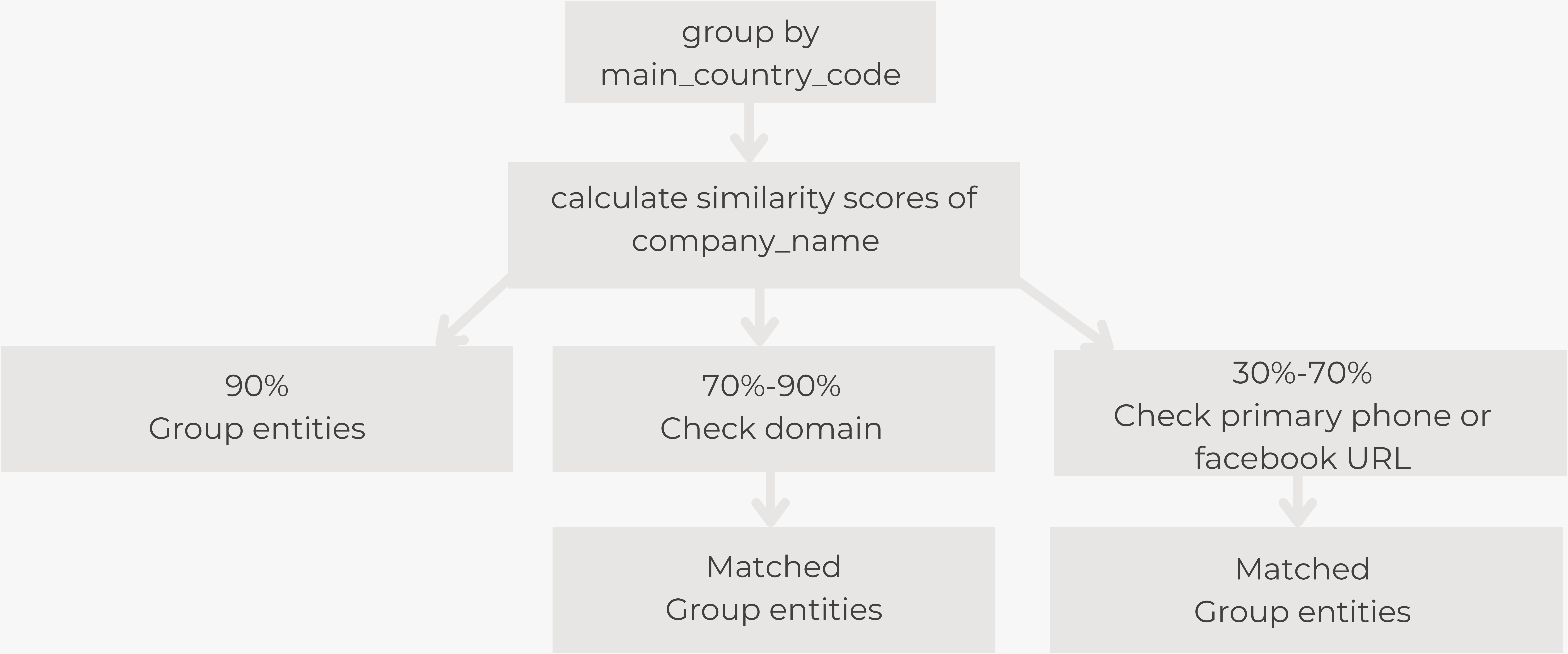
This step is essential before grouping. After early testing, where match rates were low, I added more normalization techniques beyond the basics (e.g., converting to lowercase, trimming spaces) for company names, like:

- Removed elements after the pipe symbol (|), as many values included extra info after the company name
- Limited strings to 50 characters – some entries included full descriptions rather than just names, which affected matching

Country code initially with 6.07% missing, reduced to 5.37% by extrapolating from phone number prefixes.

For website domain I eliminated the TLD, it was not necessary, because we will first group by country.

STRATEGY



group by main_country_code

calculate similarity scores of company_name

90%
Group entities

70%-90%
Check domain

Matched
Group entities

30%-70%
Check primary phone or facebook URL

Matched
Group entities

THOUGHT PROCESS BEHIND GROUPING METHOD

I chose to base the initial grouping on country, as it is the most reliable attribute and can later support the validation of subsequent groupings.

Name-based grouping is at the core of the process and includes three key distinctions. For similarity scores in the 70–90 range, I used domain comparison, as the name alone is not sufficient for a confident match. For scores between 30–70, a stronger level of validation was required — such as matching phone numbers or Facebook URLs.

For the 5% of records missing country information, even when cross-referencing multiple attributes (such as company name, domain, or email), it is difficult to establish a reliable match.

International companies often share similar attributes across different countries, which increases the risk of false positives despite these additional checks.

IMPROVEMENTS (NEXT STEPS)

The **2.48%** of records missing company names but with strong data completeness can be treated separately.

These can be matched based on domain and verified using other strong identifiers like phone number or email.

Extrapolating country from TLD can be a next good step, into reducing the **5%** of the values that aren't grouped. The downfall is that the most of them are ".com".

Automated testing on a part of the dataset to determine precision using **output values** vs **expected values** mechanism.

CONTACT

☎

0749 612 565

✉

mariamadalinaolaru26@gmail.com

in

www.linkedin.com/in/maria-mădălina-olaru-8650a318b