CleanSheet Matching Engine - Full Specification

1. Introduction

This document specifies the complete logic and matching rules required for the CleanSheet Matching Engine. The purpose of this engine is to standardize inconsistent product names from multiple files, group relevant data (such as sales and inventory), and output a unified, clean table along with a summary of match logic.

Each step must be followed with no assumptions - this document is written to avoid ambiguity for implementation.

2. Matching Logic (Step-by-Step)

2. Matching Logic (Step-by-Step)

a. Tokenization and Comparison
- Break product names into tokens: brand, size, model, variant, and features.
- Examples of tokens:
  * "Samsung TV 32 Smart" -> ['Samsung', 'TV', '32', 'Smart']
  * "iPhone 13 Pro" -> ['iPhone', '13', 'Pro']

b. Match Products by Shared Tokens
- Standardize product names where tokens clearly match.
  * "Samsung Smart TV 32" = "Smart 32in Samsung" = "Samsung 32 Smart TV"

c. Protect Against Conflicts
- If two names contain opposing tokens like "Mini" vs "Pro", they must NOT be grouped.

d. Unspecified Variant Handling
- If a vague name like "iPhone 13" is matched with multiple specific variants ("Pro", "Mini", "Gold"), mark it as:
  -> "iPhone 13 (Unspecified Variant)"
- Do not auto-group if ambiguity exists.

e. Quantity/Size as Separators
- Size/units can vary in position: "32in", "330ml", "2L", etc.
- These determine unique product types - do not group different sizes.

f. Manual Review Flags
- If a product name could match multiple candidates (same token score):
  -> Mark as "Manual Review Required"

g. Glossary-Free Matching
- No fixed glossary. Must rely on smart token logic - e.g., don't need dictionary mapping.

h. Grouping Output
- Combine cleaned, matched names from sales and inventory.
- Show all fields in one output: standardized name, total sales, total units.

i. Summary Output Table
- Must display:

CleanSheet Matching Engine - Full Specification

* Original messy names
* What each was mapped to
* Reason/score or flag status

3. Sample Inventory File

Product	Inventory Units
Samsung TV 32in S	91
Sam TV 32in	21
Samsung TV	42
Samsung TV Smart	42
Samsung 32in TV	70

4. Sample Sales File

Product	Sales (£)
Samsung TV 32in S	459
Sam TV 32in	876
Samsung TV	831
Samsung TV Smart	990
Samsung 32in TV	788

5. Sample Output: Matched Clean Table

Standardized Name	Sales (£)	Inventory Units
Apple iPhone 13 (Uns	9442	1253
Apple iPhone 13 Mini	2741	186
Apple iPhone 13 Pro	2008	271
Coca-Cola 1L Vanilla	5043	428
Coca-Cola 1L Zero	4100	444

6. Sample Output: Summary Table

Product	Standardized Name	Confidence
Samsung TV 32in S	Samsung TV 32in Smar	nan
Sam TV 32in	Samsung TV 32in (Uns	nan
Samsung TV	Samsung TV 32in (Uns	nan
Samsung TV Smart	Samsung TV 32in Smar	nan
Samsung 32in TV	Samsung TV 32in Smar	nan