SUMMARY OF INSIGHTS DRAWN

- Category listings significantly overstate catalog size. Although 100 product-category entries were scraped, deduplication revealed only 95 unique products, confirming extensive cross-listing across categories.
- Deterministic, DOM-anchored scraping proved more scalable and reliable than heuristic text parsing, enabling confident expansion after validating a small number of pages.
- Progressive sampling accelerated iteration. Reducing the working set early enabled faster validation, error correction, and refinement before final selection.
- Oversampling followed by confidence-based filtering substantially improved enrichment quality compared to enriching a fixed sample directly.
- Official brand domains consistently provided the highest-signal information for product identity and ingredients, while marketplaces were less reliable for authoritative data.
- Barcode and SKU data emerged as inherently low-signal fields. Even among major retailers, availability and consistency were limited, leading to persistently low confidence scores in early iterations.
- Ingredient validation was most effective using overlap-based scoring rather than exact matching, reflecting INCI naming variations and partial disclosures across sources.
- Strict null handling preserved dataset integrity by avoiding false precision when external confirmation was unavailable.
- Iterative tightening of validation rules improved final outcomes, with low-confidence fields identified early and excluded from the final top-10 enriched dataset.

Key Term: DOM (Document Object Model)