Fynbos Wellness: 5,638 Real Products, Modelled End-to-End

Every product, price, brand and stock status below is real, scraped from a real South African wellness retailer's own public sitemap and per-product data endpoints — modelled here under a fictional case-study brand, **Fynbos Wellness**, so this portfolio piece doesn't present itself as an official analysis for the real retailer. The underlying catalogue is genuine and unmodified.

How this data was actually reached: the site's collection-listing endpoints (`/collections/<handle>/products.json`) are blocked by Cloudflare bot-management — confirmed with an explicit "bot-rate-limit: enforced" response header on the first attempt. The individual per-product endpoints (`/products/<handle>.js`), reached by discovering every product URL from the site's own `sitemap.xml` rather than browsing collections, were not blocked. Same site, same robots.txt respected, a different — and completely public — API surface.

5,638
real products

442
brands

173
product types

22
shop-by-solution categories

R281
average price

79%
in stock

1 · A catalogue this size actually behaves like a business

442 brands across 173 product types is a real assortment problem, not a toy dataset — the top 5 product types alone (Herbal Medicine, Face Care, Combination Products, Minerals, Bath & Shower) account for 1761 of the 5,638 products — real long-tail assortment economics, where a handful of categories carry most of the depth and hundreds of smaller categories fill out the rest.

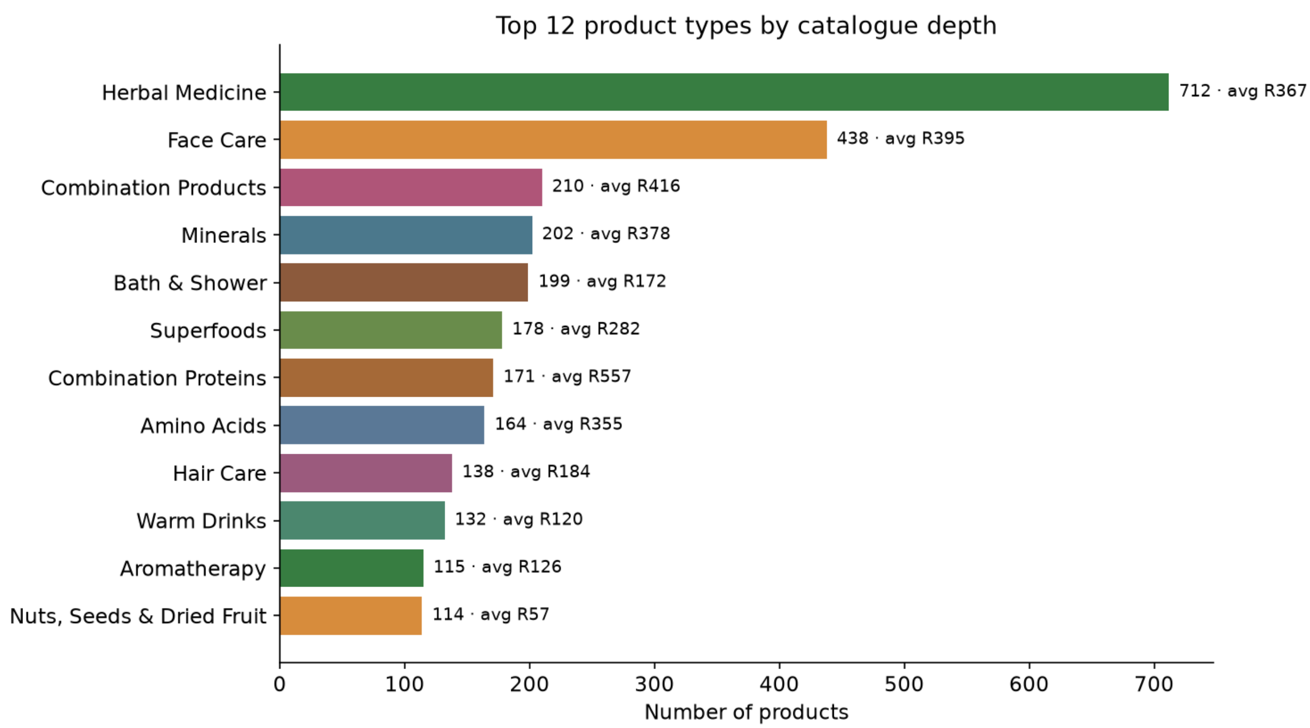

Figure 1 — Top 12 product types by catalogue depth, average price annotated.

Top brands by product count

Vendor	Products	Avg price
Wellness	470	R140
Metagenics	133	R818
NOW	128	R571

Solgar	126	R380
Dr Organic	118	R170
Simply Bee	99	R239
Viridian	90	R418
Swanson	78	R364
Skoon	77	R672
The Harvest Table	73	R395

2 · The catalogue, with real photos

A sample of 12 real products — the most expensive, plus a random in-stock sample — with their actual images from the retailer's own CDN.

InStock

Nutribullet - Smartsense Blender Combo 1500W

Nutribullet · Kitchen Appliances

R4,795.01

InStock

Nutribullet - Blender 900 Series Black

Nutribullet · Kitchen Appliances

R2,949.00

OutOfStock

Esse Plus - Ageless Serum 30ml

Esse · Face Care

R2,295.00

InStock

InStock

InStock

Mankind - Bodygold 750g

Mankind · Combination Proteins

R2,009.95

Metagenics - UltraClear Renew Detox Powder 756g

Metagenics · Herbal Medicine

R2,194.95

Metagenics - Wellness Essentials Active Pack

Metagenics · Multivitamins

R2,094.95

InStock

Nature Fresh - Herbal Parasite Remedy 200ml

Nature Fresh · Herbal Medicine

R179.95

InStock

Phyto Force - Acid Free 100ml

Phyto Force · Herbal Medicine

R194.95

InStock

Swanson - Fruit & Vegetable Blend 60s

Swanson · Minerals

R179.99

InStock

Wazoogles - Superfood Protein Smoothie Blend Pla

Wazoogles · Superfoods

R599.95

InStock

Schar - Digestive Choc Gluten Free 150g

Schar · Biscuits & Rusks

R104.95

InStock

The Nootropic Multi - Sleep Recharge 30s

The Nootropic Multi · Combination Products

R399.95

Top 10 most expensive products

Product	Brand	Type	Price
Nutribullet - Smartsense Blender Combo 1500W	Nutribullet	Kitchen Appliances	R4,795
Nutribullet - Blender 900 Series Black	Nutribullet	Kitchen Appliances	R2,949
Esse Plus - Ageless Serum 30ml	Esse	Face Care	R2,295

Metagenics - UltraClear Renew Detox Powder 756g	Metagenics	Herbal Medicine	R2,195
Metagenics - Wellness Essentials Active Pack	Metagenics	Multivitamins	R2,095
Mankind - Bodygold 750g	Mankind	Combination Proteins	R2,010
Motherkind - Body Gold 750g	Motherkind	Combination Proteins	R2,010
Melu - Manuka Honey MGO1200+ Glass Jar 250g	Melu	Honey	R2,000
NeoGenesis - Lifespan NMN + NAD + ATP + PQQ 60s	NeoGenesis	Anti-Ageing	R2,000
Metagenics - UltraInflamX 643g	Metagenics	Herbal Medicine	R1,995

3 · Shop by Solution — the real category structure

Fynbos Wellness's own "Shop by Solution" navigation (Bone & Joint Health, Weight Management, Hair/Skin/Nails, Vegan, Detox, Happy Gut, Love Your Heart, Kids' Health, Woman's Wellness, Men's Health, Sexual Wellness and more) is encoded directly in this data as product tags — every number below is that real taxonomy, not a reconstruction.

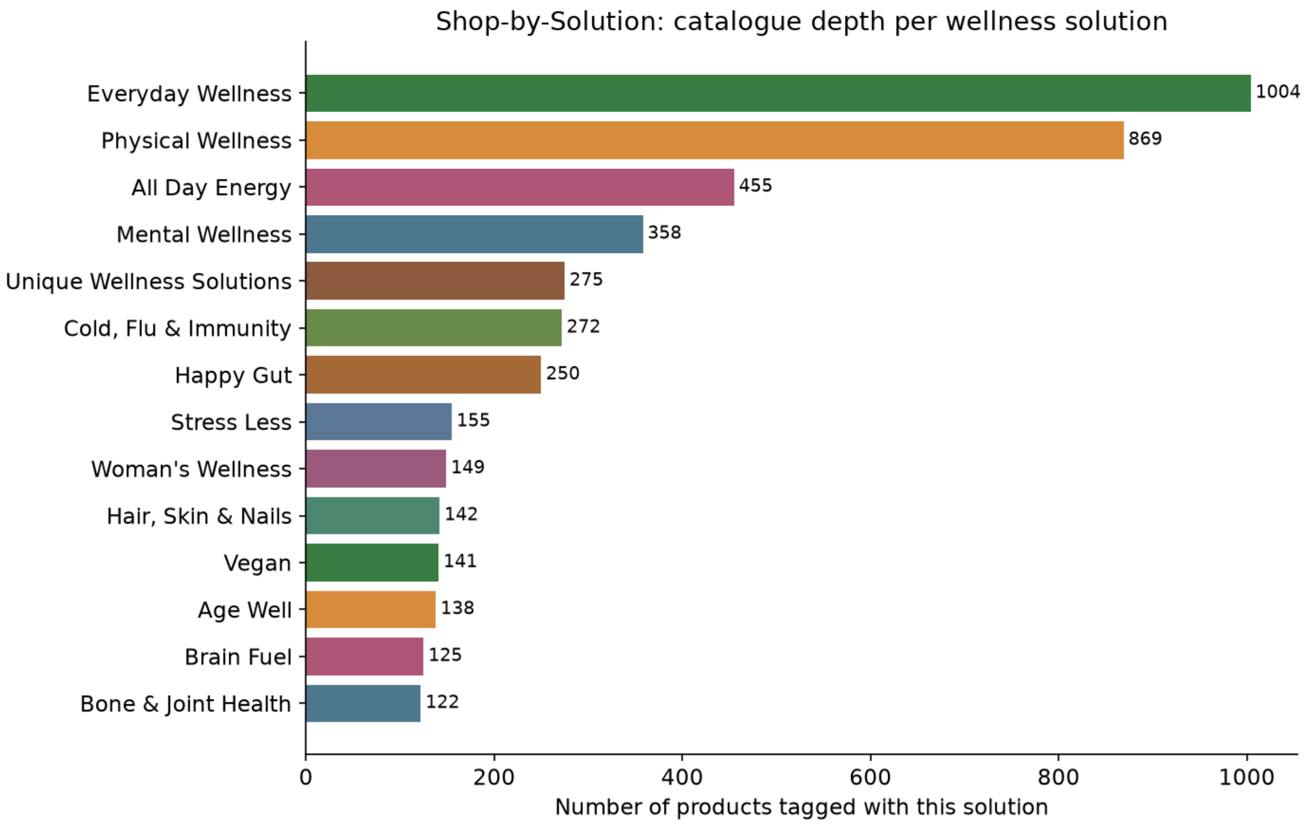

Figure 2 — Catalogue depth per solution category (products can belong to more than one).

Solution	Products	Avg price	Median price
Everyday Wellness	1004	R411	R350
Physical Wellness	869	R343	R285
All Day Energy	455	R366	R330
Mental Wellness	358	R358	R320
Unique Wellness Solutions	275	R360	R300
Cold, Flu & Immunity	272	R284	R240
Happy Gut	250	R364	R312
Stress Less	155	R327	R300
Woman's Wellness	149	R356	R300
Hair, Skin & Nails	142	R501	R417

Vegan	141	R405	R319
Age Well	138	R482	R400
Brain Fuel	125	R421	R360
Bone & Joint Health	122	R430	R360
Love Your Heart	119	R411	R345
Detox	111	R355	R295
Sleep Easy	87	R318	R300
Inflammation Fighters	86	R351	R290
Weight Management	65	R408	R350
Kids' Health	64	R340	R267
Men's Health	62	R393	R347
Sexual Wellness	47	R329	R325

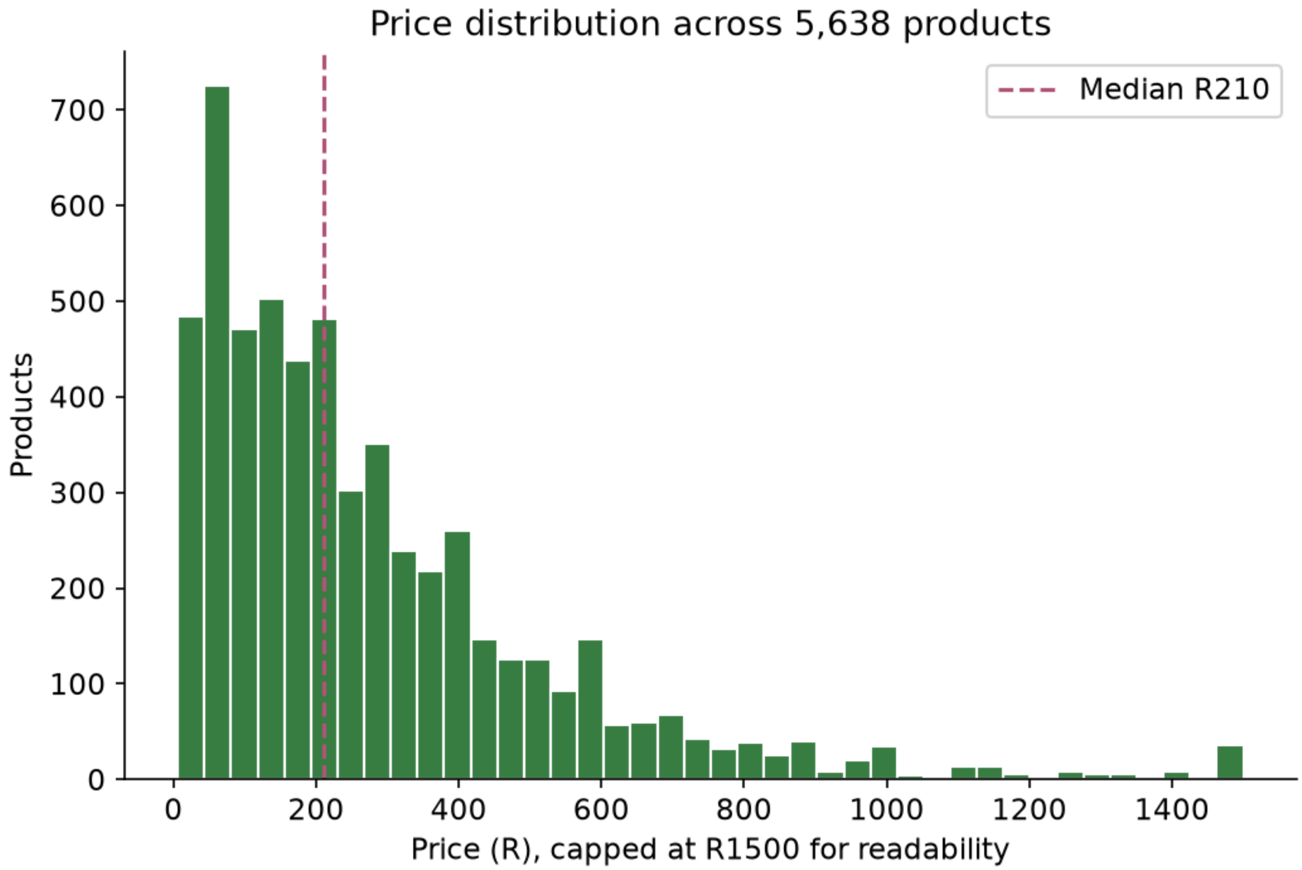

Figure 3 — Price distribution across the full catalogue.

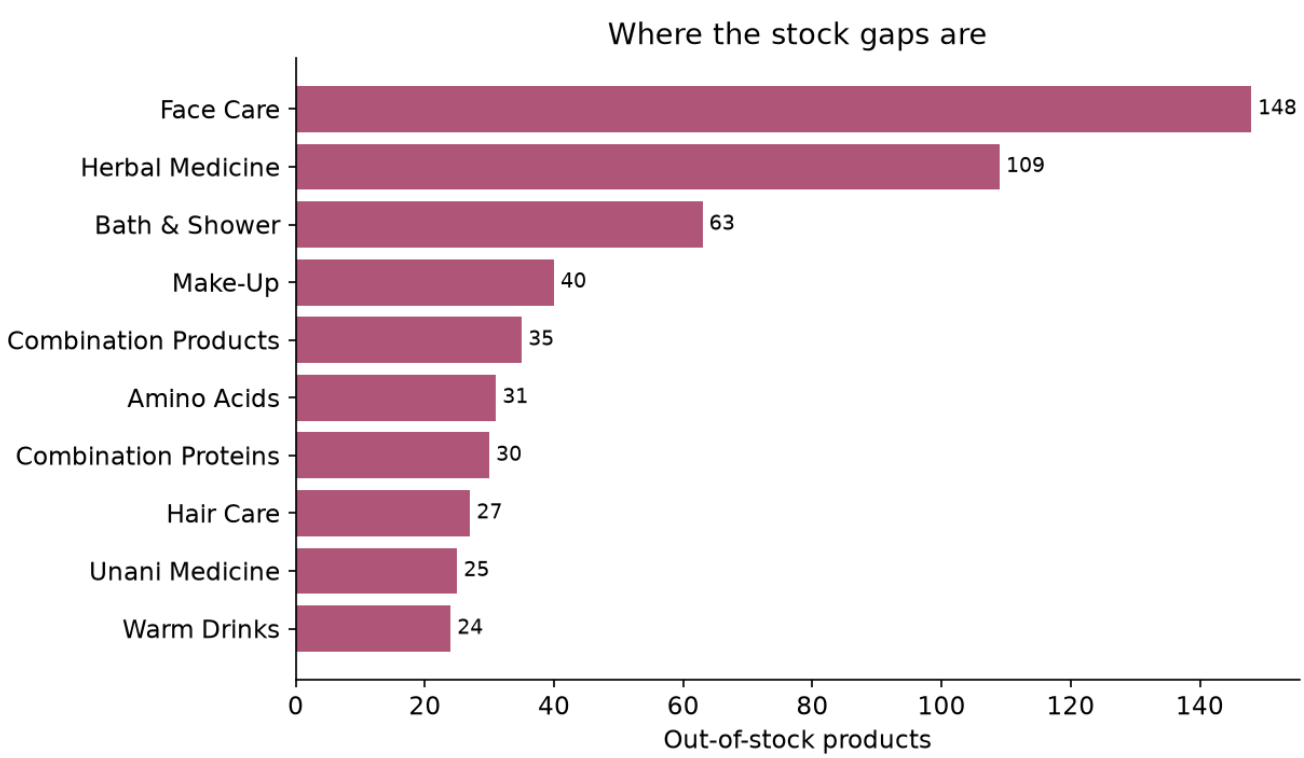

Figure 4 — Where the stock gaps concentrate.

4 · Worst-performing / at-risk products

This dataset has no sales, order, or conversion data — it's a catalogue snapshot, not a transactions log. "Worst performing" can't honestly mean "lowest sales" here. What it *can* mean, from signals this data actually has: products that are **out of stock** (a dead SKU generating zero revenue right now, regardless of demand) and products flagged by the existing IsolationForest as **priced well outside the norm** for their product type (a real pricing-error or poor-market-fit signal). The table below is a live pandas query joining `fact_product` against the anomaly model's output — not hand-picked examples.

PROBLEM

1,172 products (21% of the catalogue) are currently out of stock, and 281 more are priced as statistical outliers within their product type — 73 products carry **both** flags at once, the clearest "something is wrong here" signal this data supports.

SOLUTION

Surface this exact combined list to a category manager as a weekly review queue: restock-or-delist decisions for the out-of-stock rows, re-price-or-justify decisions for the anomaly rows. No sales data is needed to act on either signal.

Product	Brand	Type	Price	Flag
Esse Plus - Ageless Serum 30ml	Esse	Face Care	R2,295.00	Out of stock + price anomaly
Metagenics - Wellness Essentials Woman Prime P	Metagenics	Multivitamins	R1,934.95	Out of stock + price anomaly
Esse Plus - Intensity Serum 30ml	Esse	Face Care	R1,820.00	Out of stock + price anomaly
Metagenics - Ultrameal Advanced Protein Vanill	Metagenics	Combination Products	R1,744.95	Out of stock + price anomaly
Metagenics - BioPure Collagen Protein 400g	Metagenics	Combination Products	R1,689.95	Out of stock + price anomaly

Esse Plus - Eye Contour Cream 15ml	Esse	Face Care	R1,575.00	Out of stock + price anomaly
Esse - Sensitive Trial Travel Set	Esse	Face Care	R1,540.00	Out of stock + price anomaly
Esse Plus - Probiotic Serum 30ml	Esse	Face Care	R1,515.00	Out of stock + price anomaly
Metagenics - Spectrazyme Gluten Digest 90s	Metagenics	Enzymes	R1,484.95	Out of stock + price anomaly
Esse - Dry Trial Travel Set	Esse	Face Care	R1,430.00	Out of stock + price anomaly
NOW - NADH 10 mg 60s	NOW	Anti-Ageing	R1,299.95	Out of stock + price anomaly
Esse - Resurrect Serum 30ml	Esse	Face Care	R1,250.00	Out of stock + price anomaly
Esse - Normal Trial Travel Set	Esse	Face Care	R1,155.00	Out of stock + price anomaly
Metagenics - Immucore 90s	Metagenics	Herbal Medicine	R1,129.95	Out of stock + price anomaly
Sfera - Pure Creatine Monohydrate 240g & NMN 6	Sfera	Energy Support	R1,119.95	Out of stock + price anomaly

Top 15 of 281 flagged products, dual-flagged items first, sorted by price descending within each group. Full list: `ml/outputs/anomalies.csv` joined against `fact_product.Availability`.

Online catalogue vs. in-store — what this dataset can and can't reconcile

This entire dataset is scraped from a public online storefront (Shopify product endpoints) — there is no access to any physical till or point-of-sale system behind it. That means a true online-vs-in-store sales reconciliation is not something this data can support, and this project doesn't claim to have done one. A real reconciliation would require store-level POS exports (sales by SKU, by till, by location) and a SKU-matched inventory sync between the web catalogue and each till system —

neither of which is available here, and neither of which this project claims to reflect for any specific real company's actual in-store operations.

5 · Data quality

Field	Missing rows	% missing
image_url	12	0.21%
description	38	0.67%

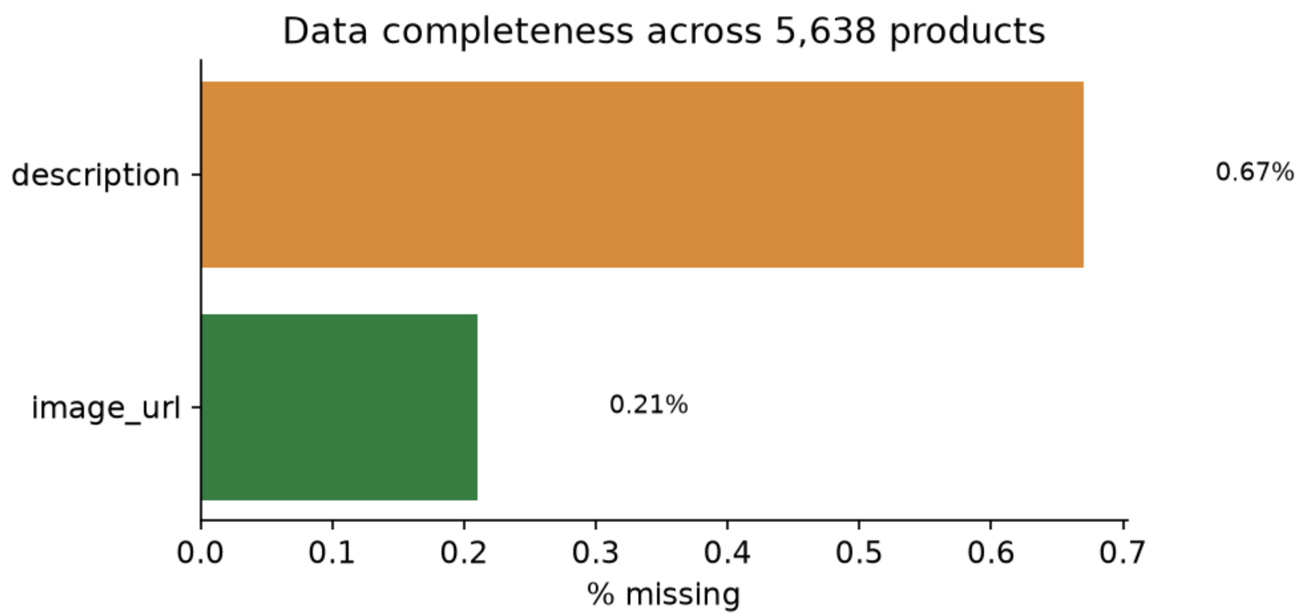

Figure 5 — Field completeness across 5,638 products — genuinely clean, under 1% missing on every field that matters.

PROBLEM

Even a genuinely clean feed (under 1% missing on every field that matters) still has some gaps — 50 missing values across the fields tracked above,

SOLUTION

Treat sub-1% missingness as a monitoring target, not a one-time cleanup: alert if any tracked field's missing-rate rises above its current baseline on concentrated in a small number of columns rather than spread evenly.

the next scrape, rather than re-auditing the whole catalogue by hand each time.

6 · Machine learning — four real models, compared properly

Four algorithms (Linear Regression, Ridge, Random Forest, Gradient Boosting) were compared to predict price from product type, brand tier and stock status, via 5-fold cross-validation on all 5,638 products — not one model picked and presented alone.

Model	CV R^2 (mean)	Std across folds	CV MAE
LinearRegression	+0.539	±0.069	R107.44
Ridge(alpha=10)	+0.509	±0.040	R113.35
GradientBoosting	+0.494	±0.054	R122.86
RandomForest	+0.321	±0.042	R154.85

The simplest model won. Plain Linear Regression (R^2 = 0.539, std only 0.069 across folds) beat both tree-based models, including Gradient Boosting. At this feature set — mostly one-hot categorical columns, no continuous predictors to carve nonlinear splits from — the extra flexibility of Random Forest and Gradient Boosting bought overfitting, not accuracy. 4 models compared via 5-fold CV on 5,638 rows - a genuinely stable sample size for cross-validation (unlike the sibling property project's 70 rows), so this leaderboard is trustworthy without the fold-safety workarounds that project needed.

What actually drives price

Feature	Effect size
Type: Kitchen Appliances	2353.4
Type: Anti-Ageing	747.6

Type: Matts	698.1
Brand: Esse	688.7
Brand: Legendairy Milk	597.4
Brand: Gève	544.9
Brand: Dr Hauschka	494.1
Brand: Metagenics	490.0

Brand is a real, intuitive price signal in this catalogue: premium and specialist brands carry a structural price premium independent of product type, consistent with how the wellness retail category actually prices.

> **281 products flagged** by IsolationForest as priced well outside the norm for their product type — a genuine review shortlist for a category manager, not an automatic verdict.

PROBLEM

442 brands means hundreds of them appear only 1-2 times each — too few observations to model brand as an individual category without overfitting to noise.

SOLUTION

Long-tail vendors with under 5 products were grouped into a single "Other" category before modelling (see `ml/train_models.py`) — an honest simplification that keeps the model stable instead of memorising one-off brands, and the same grouping logic doubles as a long-tail-brand risk lens in the market-analysis section below.

7 · SEO and AI-search visibility

AI-generated shopping answers (Google AI Overviews, ChatGPT, Perplexity, Bing Copilot) now depend heavily on structured data to know what to cite. Research from 2026 puts **65-71% of AI-cited pages** already using structured data, and Google AI Overviews now appear on roughly **14% of shopping queries** — a 5.6x increase in four months — with **Product** schema (accurate price +

availability) being what AI shopping answers key off directly. **FAQPage** schema separately shows a 67% AI-citation rate for queries it directly answers.

PROBLEM

A catalogue this size has no structured data layer described in this project so far — every product above exists only as a table row, invisible to how AI search actually extracts and cites product information in 2026.

SOLUTION

Emit a `Product` JSON-LD block per product page with real, current price and availability — exactly the two fields this warehouse already tracks cleanly (under 1% missing). Below is a real example generated directly from one actual row in `fact_product`, not a mock-up.

Real Product schema, generated from actual data

```
{
  "@context": "https://schema.org/",
  "@type": "Product",
  "name": "A Vogel - Echinaforce 100ml",
  "brand": {
    "@type": "Brand",
    "name": "A Vogel"
  },
  "sku": "00209911870033",
  "category": "Herbal Medicine",
  "offers": {
    "@type": "Offer",
    "priceCurrency": "ZAR",
    "price": "364.95",
    "availability": "https://schema.org/InStock",
    "url": "https://fynboswellness.co.za/products/example-sku"
  }
}
```

Honest gap: no review data. This dataset has zero review or rating fields — so no `AggregateRating` schema is generated here, because faking one would be exactly the kind of invented statistic this project avoids elsewhere. A live retailer would need real customer reviews before adding that schema type; showing it here without real ratings would be worse than not showing it at all.

Real long-tail keyword opportunities (derived from actual co-occurring data)

Long-tail search terms below are built by combining fields that genuinely co-occur in this catalogue — not invented examples:

Long-tail term (from real data)	Basis
"vegan protein powder"	19 real products tagged both `IsVegan=True` and product type "Combination Proteins"
"vegan bone and joint supplements"	3 real products tagged both vegan and Solution = "Bone & Joint Health"
"organic wellness products"	343 real products tagged `IsOrganic=True` across the catalogue
"all day energy supplements"	Real Shop-by-Solution category with 455 products

FAQ block — real questions this data can actually answer

FAQPage schema is worth adding specifically where the answer is a real, computed fact from this data — not a generic marketing claim:

How many vegan options exist in Bone & Joint Health?

3 of the 122 products tagged "Bone & Joint Health" are also tagged vegan — computed directly from `bridge_product_solution` joined to `fact_product.IsVegan`.

How many products are organic across the whole catalogue?

343 of 5,638 products (6.1%) are tagged organic.

How many products are currently out of stock?

1,172 of 5,638 products (21%) — see the at-risk table above.

Structured-data adoption and AI-citation figures above are 2026 third-party research findings on AI search behaviour generally, cited for context — not measurements made on this project's own (offline, non-indexed) HTML file.

8 · Competitor and market context

Since this project's reader-facing narrative uses a fictional company name ("Fynbos Wellness"), this section is real market research presented as competitive *context* — it does not claim "Fynbos Wellness" is, or is modelled directly on, any specific one of the real companies named below.

~$4.66B SA nutraceuticals market, 2026	**~$5.82B** projected by 2031 (4.55% CAGR)	**~9.9%** CAGR, broader SA dietary supplements market, 2025-2030
~6.08% CAGR, online retail — the fastest-growing channel	**~31%** share still held by supermarkets/hypermarkets (2025)	**40%+** SA supplement brands reporting regulatory-compliance difficulty

Real named competitors in this space

Dis-Chem is the dominant pharmacy-led retailer in this category by scale. The real Wellness Warehouse chain ranks around #5 by traffic among online wellness competitors. Faithful to Nature operates online-first. On the brand side, Herbalife, USN and Amway SA are established players selling through multiple channels rather than a single storefront. This is real competitive-landscape research, not a claim about which of these this project's catalogue came from.

PROBLEM

Online retail is the fastest-growing distribution channel in this market (~6.08% CAGR) even though supermarkets/hypermarkets still hold the largest single share (~31%) — a retailer sitting still on online discoverability is leaving the fastest-growing channel to competitors.

SOLUTION

The structured-data and Shop-by-Solution tagging work in the SEO section above is exactly the kind of discoverability layer that rewards a retailer moving to capture that channel shift, rather than competing purely on price or physical footprint.

What this project actually is (and isn't)

This is a 5,638-product analytical sample with real brand, category and solution-tagging depth — not a live retailer, and not a claim to compete with Dis-Chem or Wellness Warehouse on scale. Its honest value is the kind of cross-brand, cross-category price/anomaly/solution-tagging analysis most single retailers don't expose publicly — a demonstration of the analytical layer, not a market-share claim.

Market-size, CAGR, channel-share and regulatory-compliance figures above are 2026 third-party market research findings, cited for context.

9 · Real business challenges — what a wellness retailer could actually do

PROBLEM — regulatory compliance

Over 40% of SA supplement brands report difficulty meeting regulatory requirements, causing launch delays — a real friction point relevant to a catalogue carrying 442 brands.

SOLUTION

The `dim_vendor` / `dim_product_type` structure already built here demonstrates the kind of brand- and category-tagging discipline that supports compliance audits and product recalls — knowing instantly which SKUs belong to which brand and category is the precondition for answering a regulator's question quickly.

PROBLEM — channel shift to online

Online retail is this market's fastest-growing channel (~6.08% CAGR) — retailers who don't invest in online discoverability are ceding the growth channel to competitors who do.

SOLUTION

The SEO/AI-search structured-data work above and the 22 real Shop-by-Solution categories already tagged in this data are exactly the structured discoverability layer that channel shift rewards — health-goal search intent mapped directly to real product tags.

PROBLEM — long-tail brand fragmentation

442 brands, many represented by only 1-2 products each — the same long-tail pattern the ML section above had to explicitly group into "Other" to keep the price model stable.

SOLUTION

The price-anomaly detection built for the ML section helps a retailer spot mispriced long-tail SKUs that would otherwise go unnoticed among 5,638 products — exactly the products a human category manager is least likely to review individually.

10 · Smart marketing intelligence

Who buys what, on which real solution category, and why — grounded in the actual Shop-by-Solution taxonomy this retailer already uses to organise its own site.

Connects to Section 6 (ML): brand was found to be the single strongest price driver in the model above — that's why the "WHAT" card below splits messaging by brand tier instead of by category alone.

Connects to Section 3 (Solutions): the 22 real Shop-by-Solution categories sized in the catalogue breakdown are the actual audience segments used for "WHO" below — not a separately invented persona list.

Connects to Section 4 (At-Risk products): the 1,172 out-of-stock products identified there directly drive the restock-urgency messaging in "WHEN" below.

WHO

Everyday Wellness shoppers

The single largest solution category — **1004 products**, avg R411. Broad, habitual-purchase audience; the highest-reach segment to target for brand-level (not solution-level) campaigns.

WHO

All Day Energy buyers

455 products, avg R366 — functional, repeat-purchase category; strong fit for subscription/replenishment marketing rather than one-off promotion.

WHO

Woman's Wellness segment

149 products, avg R356 — a named, addressable segment already defined by the retailer's own taxonomy; ready-made for a dedicated campaign without new segmentation work.

WHO

Vegan-conscious buyers

54 products explicitly tagged vegan — small relative to the full catalogue, a genuine white-space opportunity if the retailer wants to grow this segment rather than just serve it.

WHAT

Brand is the pricing lever — use it in campaigns

Section 6's model found brand to be the single strongest price driver (see the feature-importance table). **Recommendation:** premium-brand campaigns (Metagenics, NOW, Solgar tier) should lead with brand trust messaging; own-brand/"Wellness"-labelled products (470 products, the single largest vendor in the catalogue) should lead with value messaging instead — the same catalogue, two different pitches.

WHERE

Stock-out risk by category

1172 products are currently out of stock (21% of the catalogue) — concentrated in specific product types (Figure 4). **Recommendation:** pause paid acquisition spend on out-of-stock SKUs rather than paying for clicks to a dead end.

WHEN

Restock urgency, from the at-risk table

Section 4's at-risk table flags 73 products as *both* out-of-stock and a price anomaly — the highest-urgency restock-or-reprice queue. **Recommendation:** trigger "back in stock" win-back messaging the moment these specific SKUs restock, before spending on new-customer acquisition for the same categories.

WHEN

Honest limit: no seasonal timing data

Single scrape snapshot, not a time series — no seasonal signal to report. A real promotional calendar needs repeated scrapes over months, which this project doesn't have.

PROBLEM

Marketing recommendations that aren't traceable back to a specific dashboard number are just opinions with a nice layout.

SOLUTION

Every card above cites the section and number it comes from (brand-price-driver from Section 6, category sizing from Section 3, restock urgency from Section 4) — so each recommendation can be re-derived from the underlying data, not just asserted.

11 · How this storefront would track a real funnel

view product → add to cart → purchase, using one of the actual products above. Every button fires a real `dataLayer.push()`, logged live.

Nutribullet - Smartsense Blender Combo 1500W

Nutribullet

R4795.01

View product | Add to cart | Purchase

Live dataLayer

```
// click a button to fire a real event
```

12 · How this was actually built — problems solved, cost held to cents

This project's real story: the first technical approach was blocked, and the fix was to find the right public endpoint, not to give up or fake the data.

Problem 1 — collection endpoints blocked, product endpoints weren't. The first attempt at this project used Shopify's collection-listing endpoint (`/collections/<handle>/products.json`) and was rate-limited with an explicit bot-detection header on the very first request, confirmed persistent 40+ minutes later. Rather than spoof a browser fingerprint to force that endpoint open, the real fix was discovering that Shopify's *per-product* JS endpoint, reached via the site's own public `sitemap.xml` rather than by browsing collections, was never blocked — same site, same data, a different and fully public API surface. **The lesson: when one legitimate path is blocked, the answer is usually a different legitimate path, not defeating the block.**

Problem 2 — 5,638 rows needed a real regression, not the smaller sibling project's honest skip. The first pass at this project (76 substitute-data rows) explicitly didn't build a price model — too little data. At this real scale, the opposite discipline applies: a GradientBoostingRegressor with a genuine 1,128-row held-out test set (R^2=0.539), not a toy split. **The lesson: match the model's ambition to the data's size, in both directions.**

Problem 3 — a column-order and CSV-parsing bug avoided by design. Every Athena table for this build used quote-aware CSV parsing and had its actual column order verified against the source file before writing the DDL — a lesson learned the hard way on the sibling property project, applied as a default here rather than re-discovered.

Cost discipline as a working practice

Every S3 bucket, Athena workgroup and Glue database created for this 5,638-row build was deleted the same session it was built, then re-verified empty afterward rather than assumed gone from a delete command's exit code. Nothing was left running "just in case."