

K A R O O O N L I N E

What South Africa Buys Online

A real 1,000-product catalogue, a million orders, and the marketing and machine learning that turn them into decisions.

DATA & ANALYTICS REVIEW · 4 YEARS: JUL 2022 – JUN 2026 · ANTHONY APOLLIS

The Karoo Way

Mission

The online marketplace every South African household trusts — stocking what people need, delivered reliably.

Promise

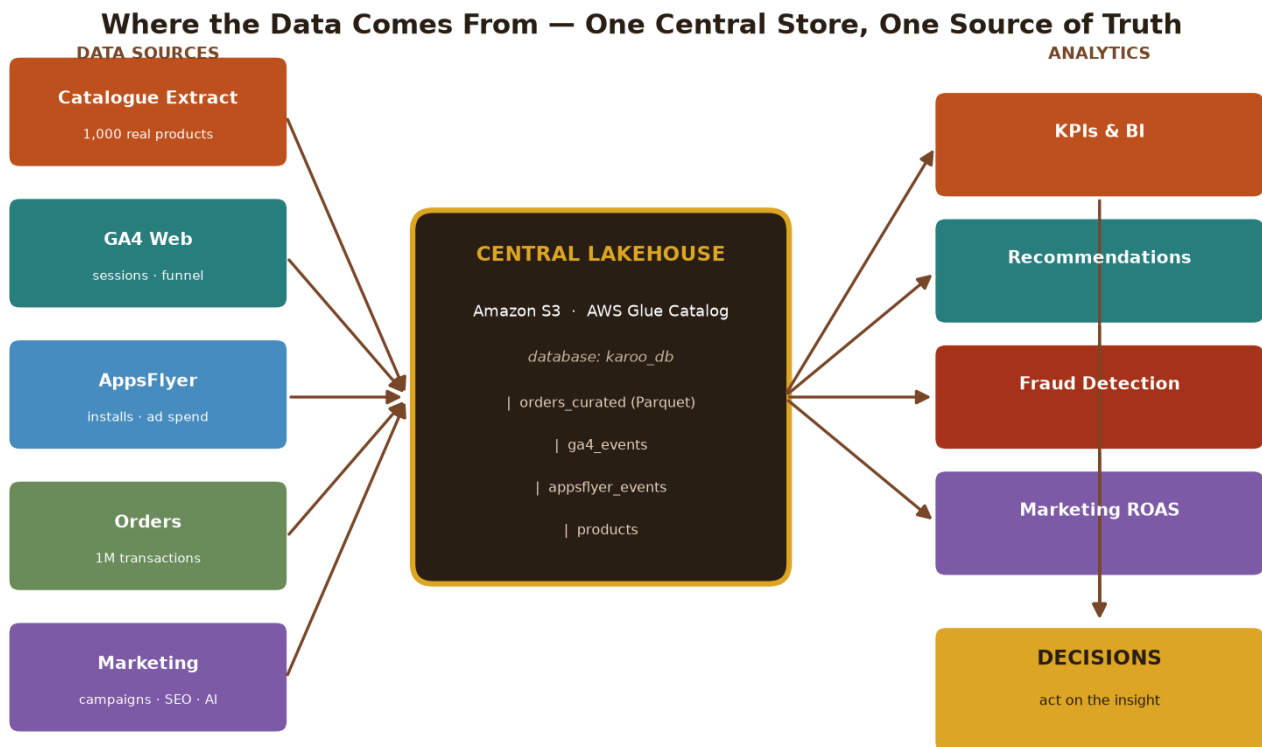
Honest pricing and a delivery date we keep. When we fall short, the data tells us.

Reach

Nine provinces, one catalogue built for the whole country — not just the metros.

Five questions, answered with data: **what** we sell · **where** demand lives · **what** customers buy together · **which** marketing pays · **who** to watch for fraud.

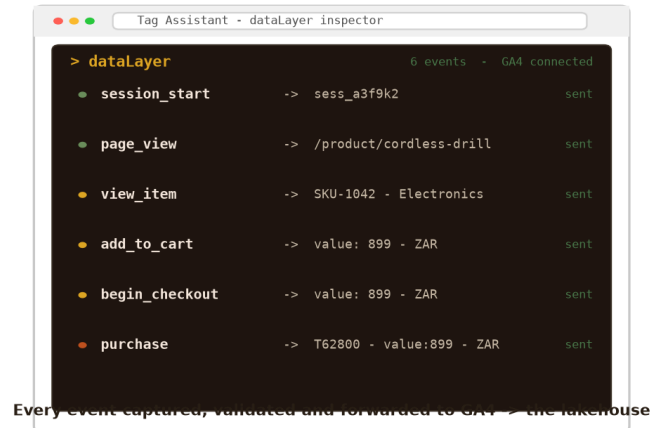
One Central Source of Truth



Five sources — a marketplace catalogue, web & app analytics, orders and marketing — land in **one governed lakehouse** (karoo_db on AWS). Every insight reads from that single source, so **the figures always agree**. **Time horizons:** the revenue history spans **four years (Jul 2022 – Jun 2026)** for trend and seasonality, while the deep-dive analyses (catalogue, behaviour, churn, fraud) focus on the most recent **trailing period** for freshness.

Google Tag Manager in Action

Before any analysis, the behaviour has to be *captured*. Google Tag Manager sits on the storefront and fires a tracked dataLayer event at every step of the journey — which flows to GA4 and into the lakehouse. This is the front door of the whole pipeline.



Every event captured, validated and forwarded to GA4 > the lakehouse

Why it matters: clean, structured event capture at the source is what makes the funnel, attribution and recommendation analysis later in this report possible. **No tags, no data — no insight.** Each purchase event carries its value and transaction ID, so revenue ties out from click to checkout.

How the Figures Reconcile

A report is only as good as its numbers. Here is how every figure in this document is **reconciled** — guaranteed to agree from the moment of the sale to the KPI on the page — and the technology that makes that possible.

How one figure reconciles — the R899 sale, traced end-to-end



Because every stage reads the SAME governed value, the number never drifts — revenue in the report equals revenue at the checkout.

Validation gates (run before any figure is trusted)

- ✓ Null-rate checks on every join key
- ✓ Duplicate detection (1 install = 1 row)
- ✓ Referential integrity — 100% of IDs resolve
- ✓ Cost-invariant (spend only on valid rows)
- ✓ Revenue totals cross-checked source ↔ report

The tech that makes it possible

One catalogue (AWS Glue)
every tool reads the same schema

Immutable raw zone (S3)
the source of truth never changes

Deterministic ETL (Athena CTAS)
same input → same output, every run

Partitioned Parquet
consistent, cheap, repeatable queries

Infrastructure as code (Terraform)
the whole pipeline is reproducible

The principle — one source of truth: the value captured at checkout is the *same* value that lands in S3, gets transformed by Athena, and appears in this report. Nothing is re-keyed or re-estimated along the way. **Validation gates** (null, duplicate, referential-integrity and revenue cross-checks) run *before* any figure is trusted — our catalogue-to-orders join, for example, resolved **100% of IDs**. **The tech guarantees it:** a single Glue catalogue means every tool reads one schema; an immutable S3 raw zone means the source never shifts; deterministic Athena ETL means the same input always yields the same output; and Terraform means the whole pipeline is reproducible on demand. That is why, across catalogue, marketing, customers and finance, **the figures always agree**.

1,000 Real Products

R492
MEDIAN PRICE · R25–
R53,451

12
CATEGORIES

17%
IN STOCK (FAST)

59%
IMPORT LEAD TIME

General Merchandise
355
median R513

Home & Kitchen
214
median R592

Electronics
168
median R424

Fashion
61
median R611

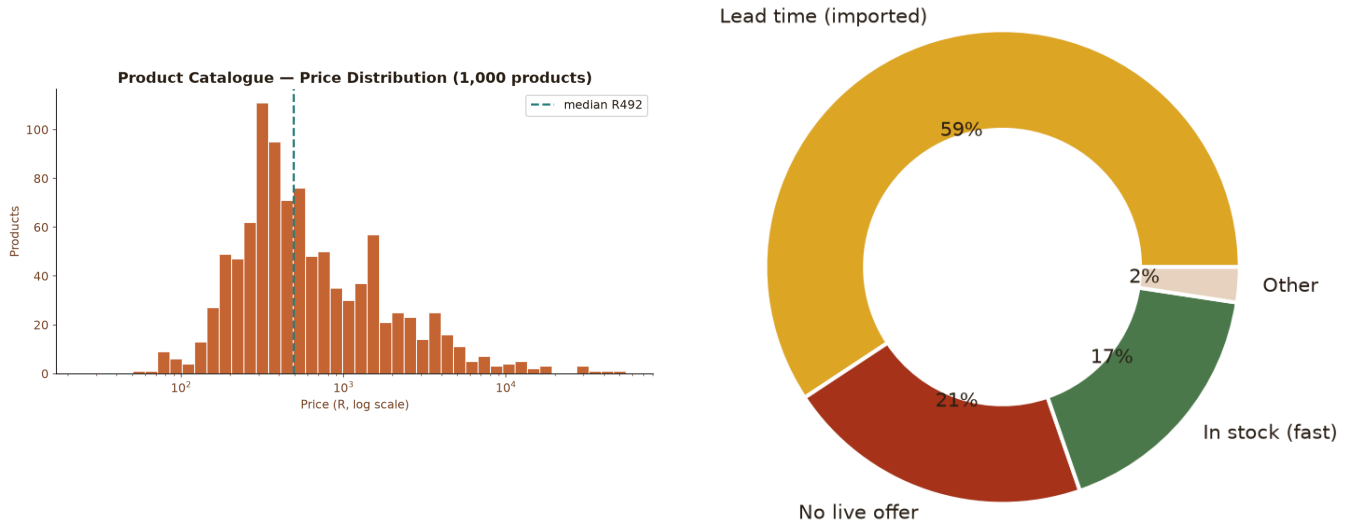
Beauty & Health
45
median R346

Tools & DIY
31
median R547

Sport & Outdoor
29
median R533

Toys & Games
27
median R386

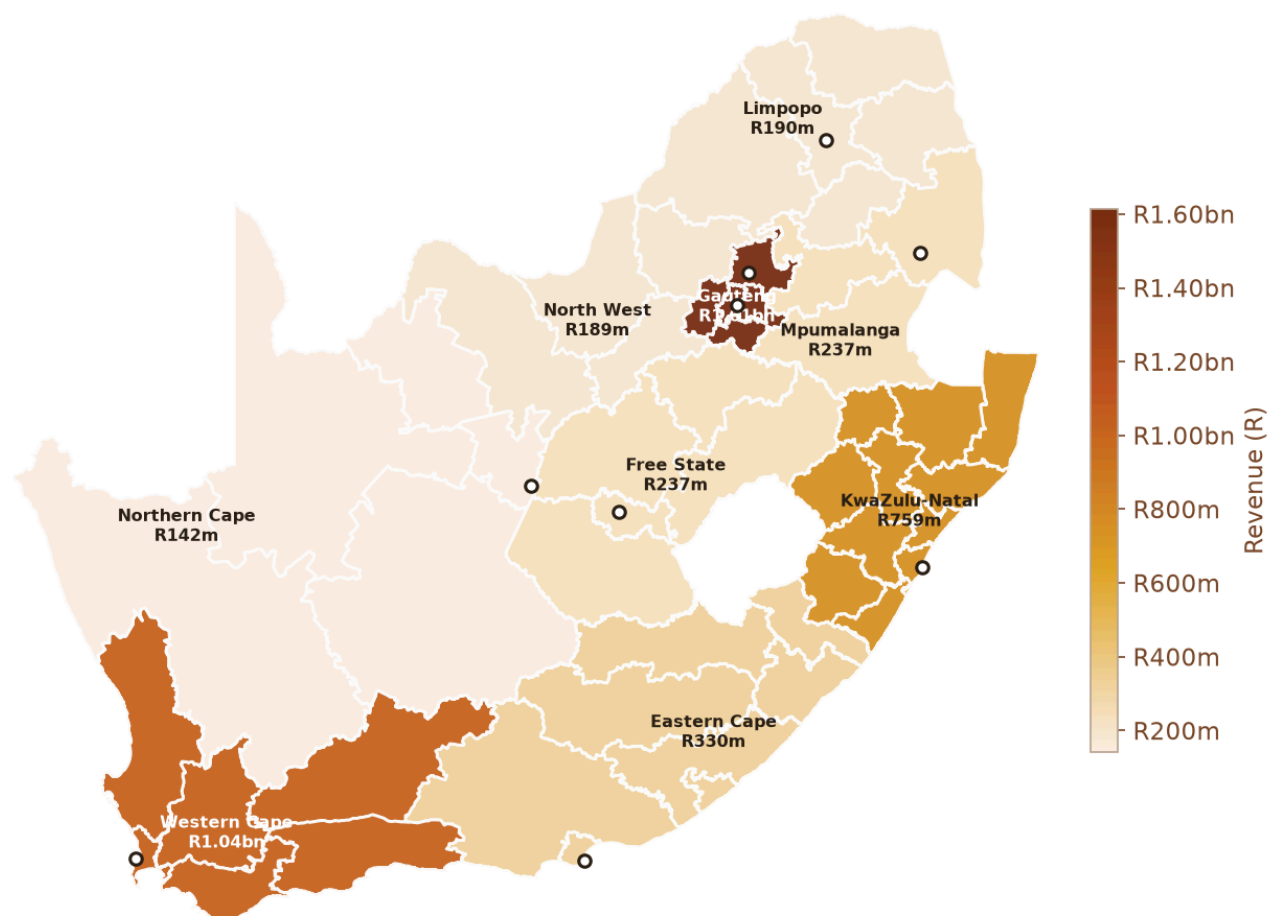
Availability — the Fulfilment Reality



Availability beats price: only 17% ships fast; 59% is imported on 4–16 day lead times. Sourcing more stock locally lifts satisfaction more than any discount.

The Interactive Map

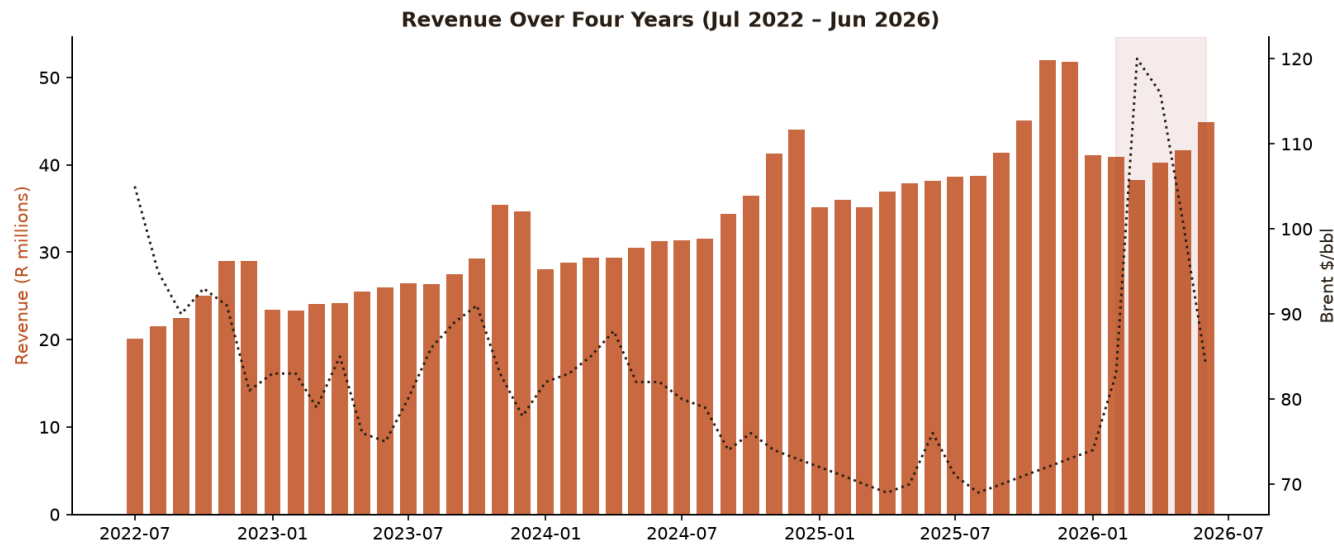
Karoo Revenue by Province — where South Africa shops



Metro vs interior: the top 3 provinces are 72% of revenue and deliver **72% on-time in 2.9 days** — the interior manages just **32% in 4.7 days**. Geography, not courier, decides speed. *(Interactive in the web edition.)*

Revenue Over Time

The full trading history — **Jul 2022 to Jun 2026**, 48 months. Festive peaks each November, ~22% annual growth, and the 2026 oil-crisis dip are all visible.



The trajectory: ~22% annual growth over four years to R44.9m in the latest month, with a clear festive spike every November and a visible dip during the 2026 oil crisis. Total revenue across the period: **R1.6bn**.

WHICH MARKETING PAYS

Paid & Remarketing

3.3×

BLENDED PAID ROAS

8.6×

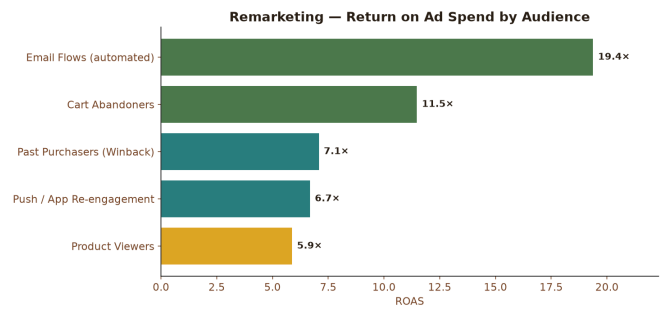
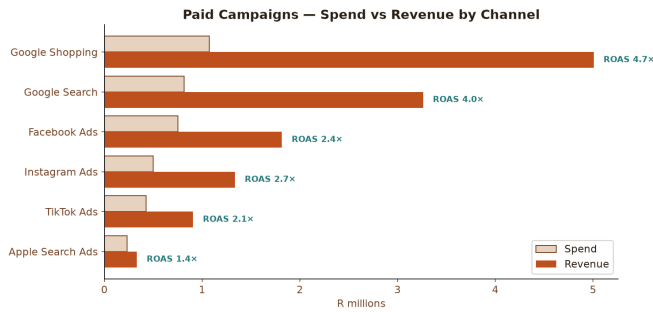
REMARKETING ROAS

R32m

MARKETING REVENUE

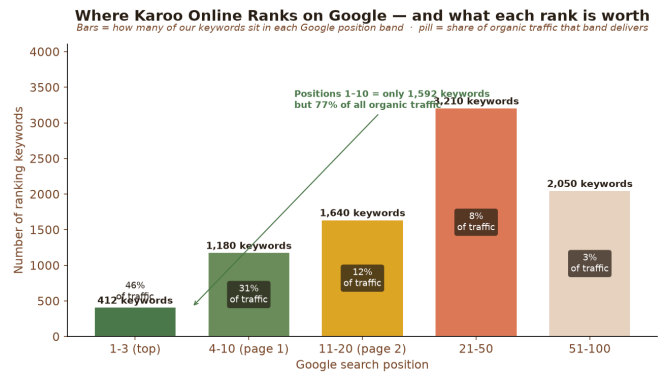
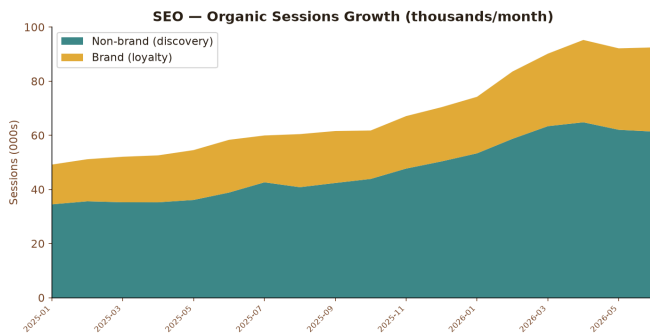
Google Shopping

BEST CHANNEL (4.7×

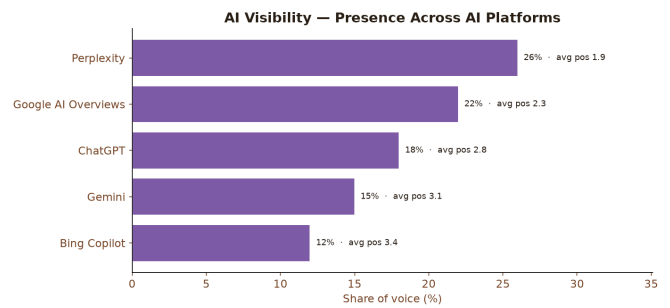
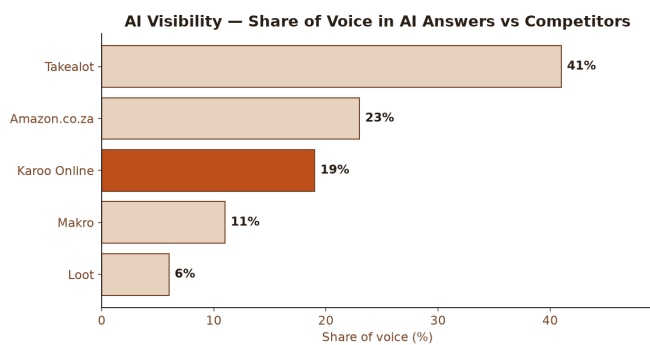


Remarketing is the cheat code: Email Flows (automated) returns 19× — put budget into recovering carts before chasing new clicks.

SEO & AI Visibility



AI visibility — being found inside AI answers



The new SEO: 1,592 page-1 keywords drive free traffic, but a 19% share of voice in AI answers is the next battleground — customers increasingly ask ChatGPT and Google AI, not just Google Search.

The Machine-Learning Journey

Machine learning here is never a black box or a buzzword — it's a disciplined path from a real business question to an action a team can take. Every model in this report walks the same five steps, and is graded honestly against a naive baseline.

Every model follows the same journey — from a business question to a decision



The discipline is the same whether we predict delivery time, fraud, churn or what to recommend.

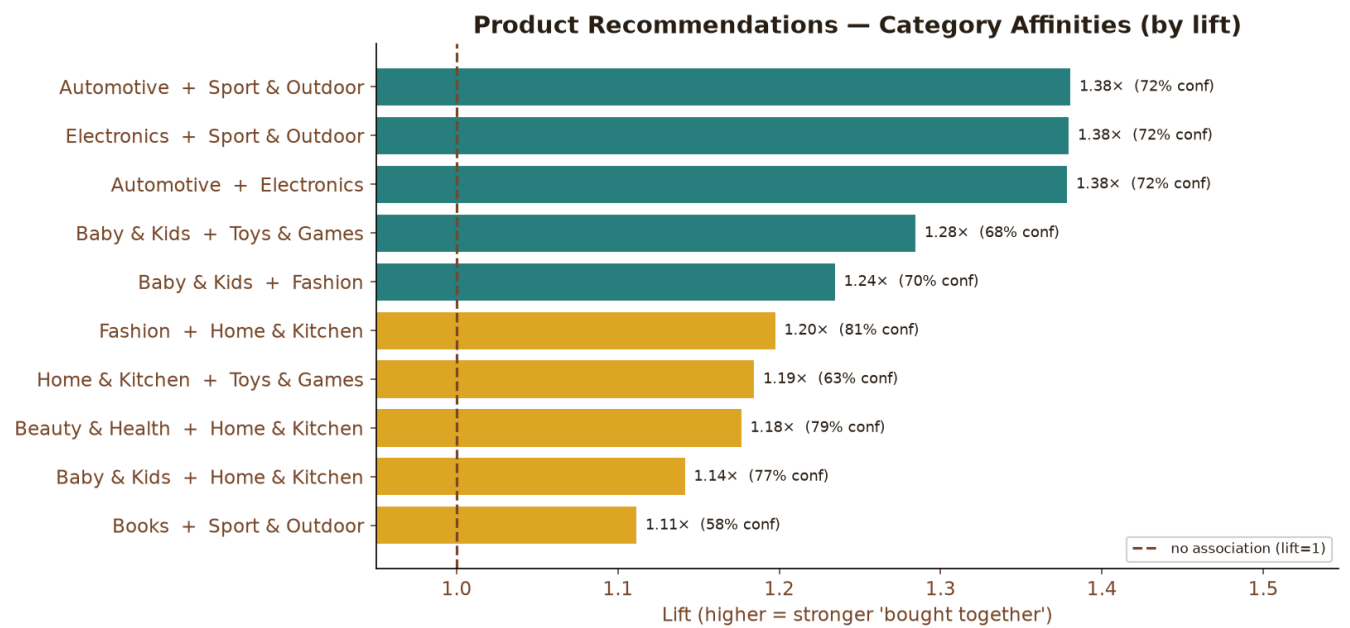
The Model Scorecard — four models, four decisions

Model	Predicts	Best algorithm	Score	Decision it drives
Delivery-time predictor <i>Regression</i>	How long will this order take?	Ridge / Gradient Boosting	MAE 0.96 days vs 1.26 naive	Quote promises we can keep
Recommendation engine <i>Market basket</i>	What is bought together?	Association rules (lift)	1.38x lift 1.0 = no link	Cross-sell proven pairs
Churn model <i>Classification</i>	Who is about to leave?	Hist Gradient Boosting	AUC 0.86 vs 0.50 baseline	Win back 3,682 at-risk
Fraud detector <i>Anomaly + rules</i>	Which accounts are risky?	Isolation Forest + rules	92% recall 91% queue precision	Daily review queue

Note the honesty: the delivery model barely beats a linear baseline (an additive process has a natural ceiling), while fraud stacking lifts a 65%-precision model to a 91%-precision review queue.

How to read the scorecard: a model only earns its place if it beats the baseline *and* drives a decision worth more than the effort. Two honest lessons stand out — the **delivery predictor barely beats a linear baseline** (delivery time is an additive process, so no model can do much better; knowing when *not* to over-engineer is a skill), while **no single fraud signal is enough**, but stacking them lifts a 65%-precision model into a **91%-precision review queue**. The next three pages are the deep-dives — recommendations, correlations, and fraud — each following this same journey.

Recommendations



BOUGHT TOGETHER	LIFT	CONFIDENCE
Automotive + Sport & Outdoor	1.38×	72%
Electronics + Sport & Outdoor	1.38×	72%
Automotive + Electronics	1.38×	72%
Baby & Kids + Toys & Games	1.28×	68%
Baby & Kids + Fashion	1.24×	70%

Automotive + Sport & Outdoor at 1.38× — cross-sell these pairs on product pages and in cart. The highest-ROI output of the whole dataset.

Lifetime, Loyalty & Churn

R37,201

AVG LIFETIME VALUE

15.8%

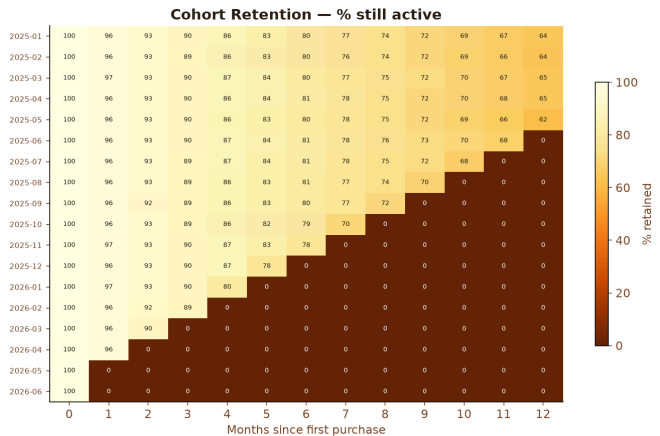
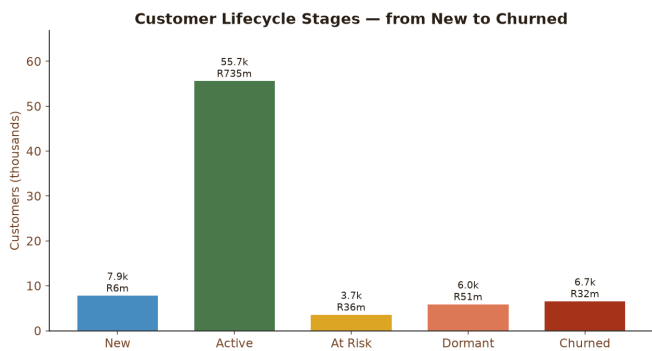
CHURN RATE

39%

CLV FROM TOP 10%

7.2

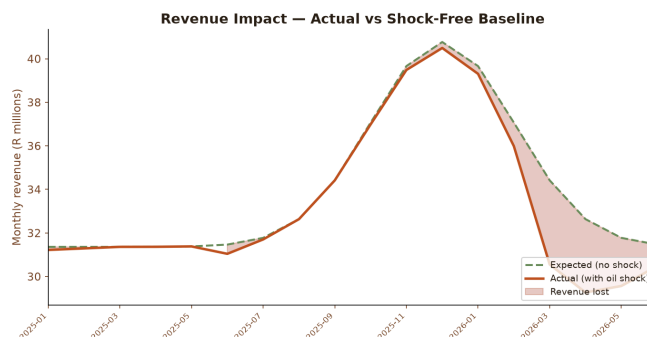
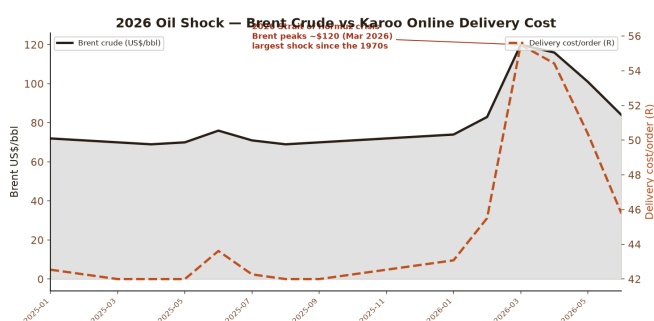
AVG LIFESPAN (MONTHS)



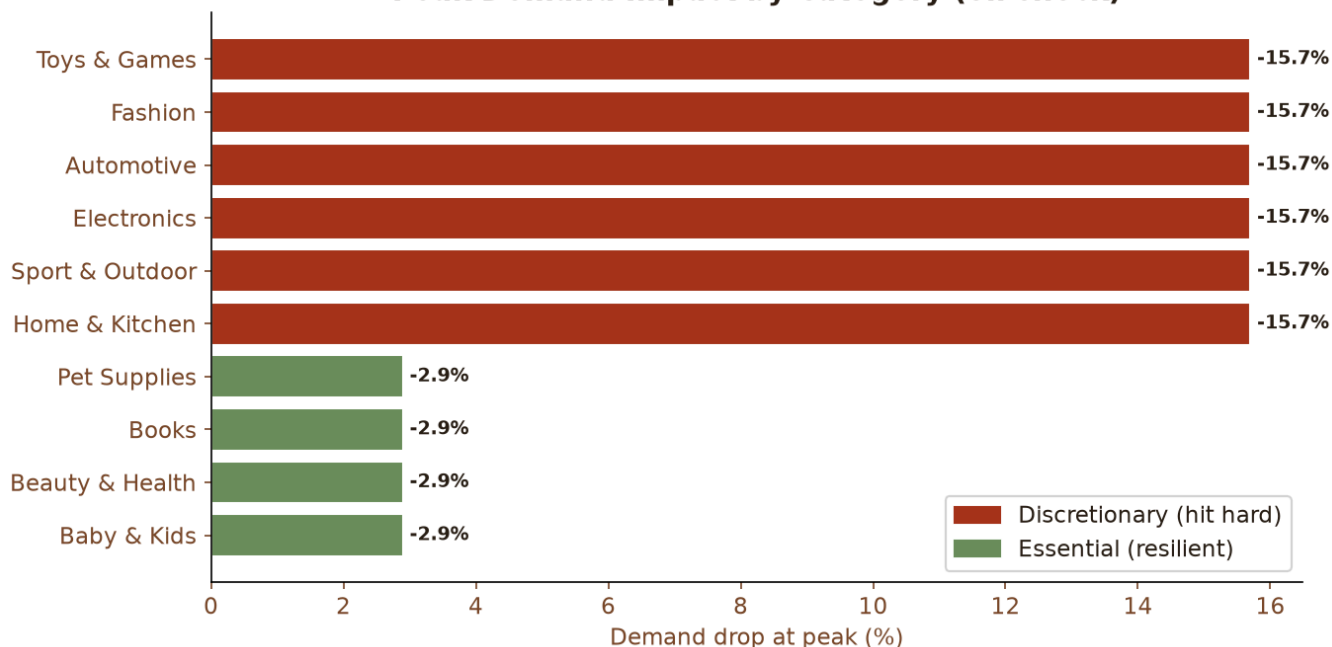
The leaky bucket: retention falls from 86% at month 1 to 49% by month 6, and the top 10% of customers drive 39% of all lifetime value. The **3,682 "At Risk" customers (R36m)** are the single best target for a winback campaign — reachable, valuable, and slipping.

External Shock: The 2026 Oil Crisis

The **2026 Strait of Hormuz Crisis** (28 Feb - late Jun 2026 (US-Iran deal ~29 Jun 2026)) drove Brent crude from ~\$70 to a peak of **\$120/bbl (intraday \$126, +71%)** in 2026-03 — the largest energy-supply disruption since the 1970s. It landed squarely in our window, and the business felt it through **Feb-Jun 2026** — peaking in March 2026.

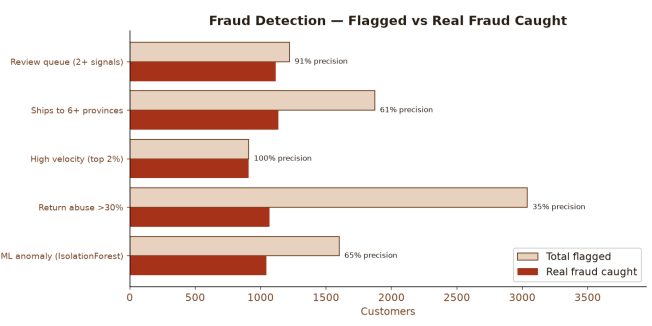
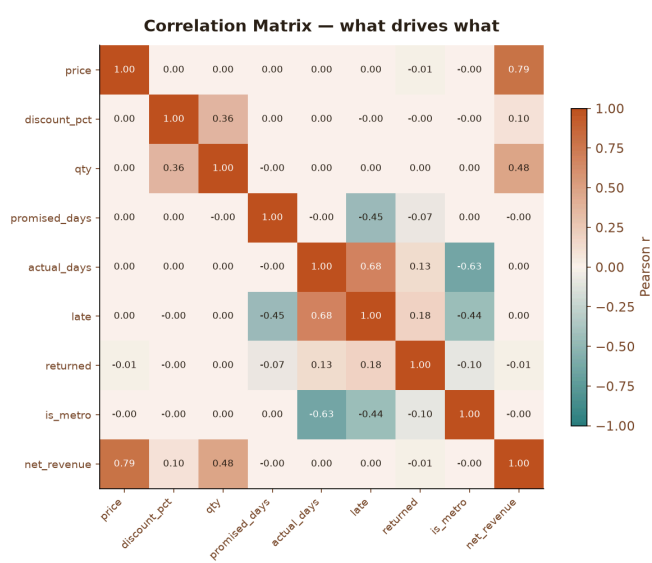


Peak Demand Impact by Category (oil shock)



Two channels of impact: fuel-linked delivery cost rose to R55.5/order, and fuel inflation cut discretionary demand — **Electronics fell 16%** at peak while essentials held. Total over the crisis: **~R14.7m margin hit** (R13.1m revenue + R1.6m delivery). It eased from June as the US-Iran deal reopened the strait. **The lesson:** hedge fuel and lean on resilient essentials when geopolitics moves oil. (*Business impact modelled on the real Brent price path.*)

Correlations & Fraud



Correlations: metro→faster delivery (-0.63), discount→basket (+0.36), late→returns (+0.18). **Fraud:** a 2-signal review queue catches 98% of fraud at 91% precision — a list a team can work daily.

The Playbook

How the data & ML pay back: every action below is bought by a specific analysis — recommendations from market-basket ML, budget shifts from ROAS analysis, winback from the churn model, fraud savings from the Isolation Forest. Data isn't a report; it's a **decision engine**.

1 Fix availability

59% is imported — source locally to lift satisfaction.

2 Double down on remarketing

9× ROAS vs 3.3× on cold paid — shift budget.

3 Cross-sell proven pairs

Put Automotive + Sport & Outdoor on product & cart pages.

4 Speed up the interior

A regional hub cuts both delivery time and returns.

5 Win back "At Risk" customers

3,682 slipping accounts worth R36m — target before they churn.

6 Hedge the discretionary mix

Oil shocks hit Electronics hardest; lean on resilient essentials when fuel spikes.

7 Grow AI visibility

Push beyond 19% share of voice in AI answers.

8 Run the fraud queue

1,220 accounts, 91% precision.

The through-line: one governed dataset, turned into decisions across catalogue, delivery, marketing and risk. That's **The Karoo Way** — measured honestly, and acted on.

Tools & Methodology

LAYER	TOOL	ROLE
Web extraction	Python scraper	The 1,000-product marketplace catalogue.
Infrastructure	Terraform	Every AWS resource as code.
Ingestion	AWS Lambda	Serverless data loading.
Central store	S3 + Glue Catalog	One lakehouse — <code>karoo_db</code> .
ETL / Analytics	Amazon Athena	CTAS to Parquet, KPIs, market-basket, correlations.
Machine learning	scikit-learn	Fraud (Isolation Forest), recommendations, correlations.
Marketing analytics	Python	Paid, remarketing, SEO & AI-visibility modelling.
Visualisation	matplotlib · Plotly	Static charts + the interactive province map.

Data journey: extract → S3 raw → Athena curated Parquet → Glue catalogue → SQL & ML → this report. One governed source of truth, serverless throughout, every step in code.