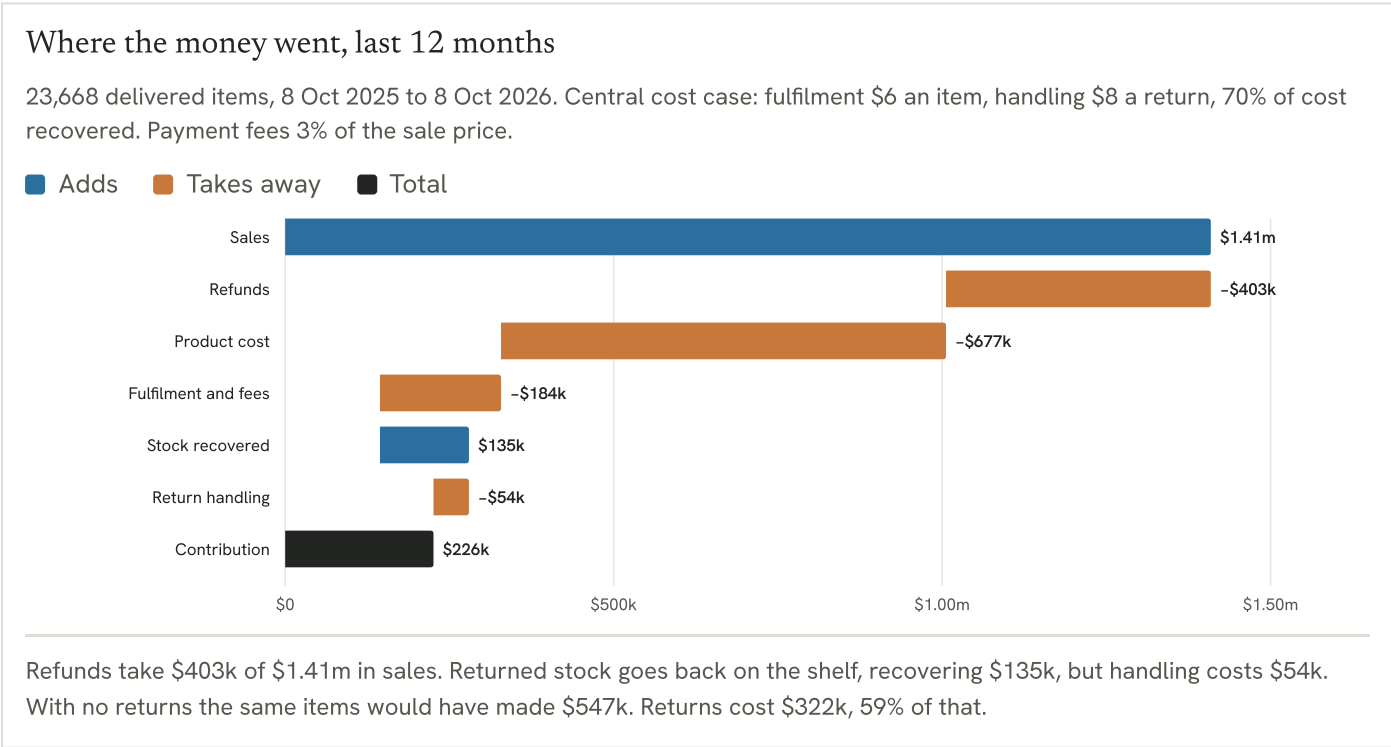


Returns cost this shop more than half its contribution. Nothing in the data says which items will come back, so fix the whole shop, not a list of "bad" products.

180,516 order items from 79,915 customers of TheLook, a clothing shop Google made up for its public data, 11 Jan 2019 to 11 Oct 2026. We checked how the data was made, built a contribution model with every cost assumption in the open, looked for any group or product that returns more than chance, and tested whether a model could predict a return when the order is placed.

Synthetic demonstration. Not a client result. The customers, orders and returns were generated by a program, so the patterns here say nothing about real shoppers. Every money figure is a scenario built on the assumptions shown.

Business: online clothing shop (made up) Extract: 2026-10-11 18:18:40 UTC
Delivered items studied: 61,042 Prepared by Dunwell Insights, October 2026



Returns are the biggest cost after the products themselves, and they look the same everywhere

28.4% Of delivered items came back range 27.9% to 28.9%	\$322k Cost of returns, last 12 months 59% of contribution, central costs	\$47 Lost on an average return refund, less stock recovered, plus handling	43 Returns a year a \$2,000 fix must prevent out of about 6,745 a year
--	--	--	---

What we found

- 1

The order statuses are made up at random.

Orders placed five years ago are still "processing" as often as orders placed this week. We kept to delivered items and treated each return as a full refund, because there is no refund ledger.
- 2

About 28% of delivered items come back, and returns cost \$322k a year

under central costs: 59% of what the same items would have made with no returns. The range across our low and high cost cases is \$237k to \$426k.
- 3

No category returns more than any other beyond chance.

Raw category rates run from 17% to 32%, but small categories swing more, and a test for any real difference finds none (adjusted p = 0.93). Brand, price, country and customer type are the same. Only order year shows a real difference, from the two earliest years.
- 4

Products that look like "serial returners" are what chance gives.

11.3% of products had every delivered item returned; a 28% coin toss gives 11.5%. No product has enough sales to judge on its own.
- 5

No model could predict a return when the order was placed.

Logistic regression and LightGBM did no better than the category rate on the sealed test period. We report that rather than add signal that isn't there.
- 6

So the cost ranking is just price times volume.

Outerwear & Coats tops it because it sells many expensive items, under every cost case we tried. A \$2,000-a-year fix there must cut its return rate by 1.5 points.
- 7

A test in one category would take too long.

Outerwear & Coats delivers about 22 items a week, so a 20% cut would take 86 weeks to show. Shop-wide, the same cut shows in about 5 weeks.

What we'd do

Test clearer size guidance across the whole shop, randomised by product. Returns cost about \$47 each, so a \$2,000-a-year fix pays for itself if it prevents 43 returns a year, a cut of 0.18 points. Run it for at least 17 weeks to see a 10% cut.

Don't single out products or customers from this data: their return rates are chance. With a real shop's data, the same checks would show whether some products truly return more, and those would go first.

Separately, 2,088 items in the last year (8.8%, all priced \$18 or less) lose money even when nobody returns them, once fulfilment and fees are paid. That is a pricing question, not a returns one.

1 · DATA CHECKS · **SYNTHETIC DATA**

Statuses are handed out at random, so we kept to items that were delivered

The extract has 181,946 items. 1,430 were dated after the extract was made and were dropped. Item and order records agree everywhere, prices never vary from the list price, and the stock cost always equals the product cost. But the dates are not those of a real shop: 34,728 items were dispatched before they were ordered, and the share of orders in each status doesn't change with age.



What we checked

Check	Result
Item status matches its order	Yes
Items per order match the order's count	Yes
Sale price equals list price (no discounts)	Yes
Product cost equals stock cost	Yes
One status per order	Yes
Every item has a product and a customer	Yes
No duplicate items	Yes
Items dispatched before they were ordered	34,728
Return after delivery	0 to 3 days
Delivery after dispatch	0 to 5 days

Who counts

Items	Count
In the extract	181,946
Dated after the extract, dropped	1,430
Cancelled, not counted	27,258
Processing, not counted	36,205
Shipped, not counted	53,871
Delivered (complete or returned)	63,182
Delivered less than 3 days ago, not yet counted	2,140
Studied	61,042
of which returned	17,345 (28.4%)
Cancelled, processing and shipped items are left out of return rates. In a real shop that could bias the rate; here the statuses are random, so it doesn't.	

2 · MONEY · **SIMULATED, ASSUMPTION-BASED**

Every return costs about \$47, and the assumptions move the total by a lot

There is no refund ledger, so a return is treated as a full refund. Fulfilment, payment fees, handling and how much of a returned item's cost is recovered are not in the data. We set them as assumptions and show three cases. Expected contribution per item = $(1 - p) \times \text{sale price} - \text{cost} - \text{fulfilment and fees} + p \times (\text{stock recovered} - \text{handling})$, where p is the chance of a return.

Last 12 months under three cost cases

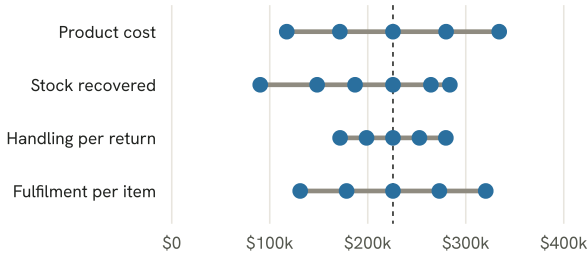
Low-cost case: fulfilment \$4 an item, handling \$4 a return, 100% of cost recovered. Central: fulfilment \$6 an item, handling \$8 a return, 70% of cost recovered. High-cost case: fulfilment \$8 an item, handling \$12 a return, 30% of cost recovered. Payment fees 3% of the sale, not refunded.

Last 12 months	Low-cost case	Central	High-cost case
Sales	\$1,409,045	\$1,409,045	\$1,409,045
Refunds on returns	-\$403,079	-\$403,079	-\$403,079
Product cost	-\$677,432	-\$677,432	-\$677,432
Fulfilment and payment fees	-\$136,943	-\$184,279	-\$231,615
Returned stock recovered	\$193,495	\$135,447	\$58,049
Return handling	-\$26,980	-\$53,960	-\$80,940
Contribution	\$358,105	\$225,741	\$74,026
Contribution if nothing came back	\$594,669	\$547,333	\$499,997
Cost of returns	\$236,564	\$321,593	\$425,971
Share of contribution lost to returns	40%	59%	85%

Product cost is paid on every item. A returned item comes back to stock and is worth the recovered share of its cost.

How much each assumption matters

Cost of returns over the last 12 months as one assumption moves and the others stay at the central case



Contribution after returns, last 12 months. The dashed line is the central case, \$225,741. Fulfilment and product cost change what the shop makes but not what returns cost. Handling and how much returned stock is worth decide the cost of a return.

Formula check

Central case	Last 12 months	All years
Bridge, step by step	\$225,741	\$600,030
Formula with actual outcomes	\$225,741	\$600,030
Difference	\$0	\$0
Formula with return rates	\$227,897	\$601,406

The formula, run with each item's actual outcome, gives the bridge total to the cent. Run with the shrunk return rates, it lands within 1%.

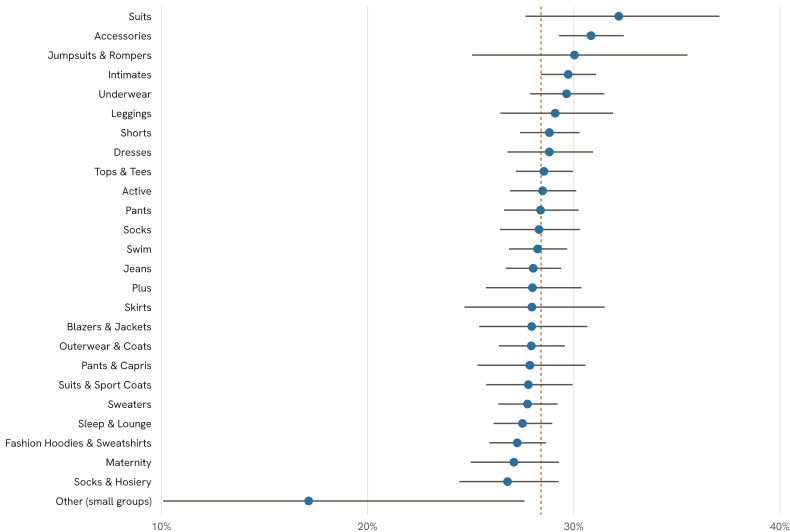
3 · RETURN RATES · SYNTHETIC DATA

28.4% of delivered items come back, and no category stands out beyond chance

We worked out the return rate for every category, brand, price band, country, distribution centre and type of customer, with a 95% range. Then we pulled each group toward the shop-wide rate by an amount set by how much the groups really differ (beta-binomial shrinkage). For every grouping but order year the data says they don't differ, so every group's steadied rate is the shop-wide rate.

Return rate by category, every year

Delivered items past the 3-day return window. Lines are 95% ranges. The dashed line is the whole shop.



The highest raw rate is 32% and the lowest 17%, but the ranges overlap the shop line almost everywhere, and the spread is what chance gives for groups this size.

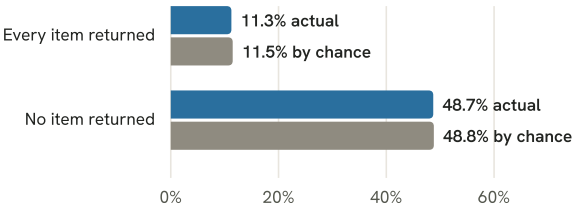
Every grouping we tried

Could the differences between groups be chance? Lower p means less likely. Adjusted for testing ten groupings at once.

Grouping	Groups	Raw range	Adjusted p
Order year	8	23.8% to 29.0%	0.010
How the customer found the shop	5	27.7% to 31.1%	0.15
Customer country	12	21.5% to 29.9%	0.22
Category	26	17.1% to 32.2%	0.93
Distribution centre	10	27.5% to 29.7%	0.93
Price band	5	27.9% to 29.2%	0.99
Items in the order	4	28.1% to 29.2%	0.99
Department	2	28.4% to 28.5%	1.00
Brand	31	23.2% to 32.1%	1.00
Customer gender	2	28.4% to 28.5%	1.00

"Serial returner" products are what chance gives

Share of products with every item returned, and with none, against what a 28% coin toss would give



The median product sold 2 delivered items in five years, and none reached the 30 we set as the minimum to judge a product on its own.

No model could tell which orders would come back

The question: when an order is placed, will this item come back? Each model saw only what was known then: category, brand, price, cost, items in the order, distribution centre, the customer's age, country, how they found the shop, time since sign-up and their earlier orders and returns. Trained on orders up to 2025, tuned on January to April 2026, and scored once on 12,946 items ordered from 1 May 2026, after the choice was frozen.

Every model on the test period

Model	Log loss	Brier	PR-AUC	ROC-AUC	Average prediction
Whole-shop rate	0.5989	0.2044	28.6%	0.500	28.3%
Category rate (the baseline)	0.5989	0.2044	28.6%	0.500	28.3%
Logistic regression	0.5989	0.2044	29.6%	0.511	28.4%
LightGBM	0.5996	0.2047	28.2%	0.500	28.6%
Actual return rate, test period	-	-	-	-	28.6%

Log loss: lower is better. The best model, logistic regression, differed from the category rate by less than 0.0001, and its range (-0.00007 to 0.00003) includes zero. ROC-AUC of 0.5 means no better than a coin toss at ranking.

Decision rule, set before the test: use a model only if it beats the category rate with the range below zero. Neither did, so the cost ranking uses the steadied group rates. In this made-up data, returns look like a 28% coin toss on every item. A real shop's data would usually show some pattern (fit-sensitive categories, sizes, repeat returners), and this is the test that would find it.

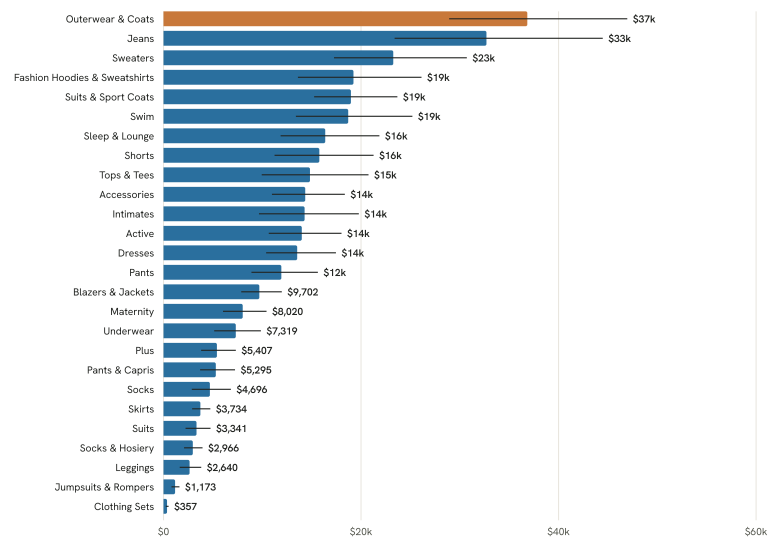
5 · WHERE RETURNS COST MOST · **SIMULATED, ASSUMPTION-BASED**

Outerwear & Coats costs the most in returns because it sells the most expensive items

Cost of returns a year = items delivered × return rate × (sale price – stock recovered + handling). With every return rate the same, the ranking follows sales: the correlation between a category's cost of returns and its sales is 0.996. The top ten is the same in 9 of 10 places under the low-cost case and 9 of 10 under the high-cost case, and Outerwear & Coats is first in all three.

Cost of returns a year, by category

Last 12 months. Bars use the central costs; the line runs from the low-cost to the high-cost case.



Cost of returns depends on handling and stock recovery, not fulfilment. The ranking hardly moves between cases, so the assumptions change the size of the problem, not where it is.

What a \$2,000-a-year fix needs to do

Category	Items a year	Average price	Loss per return	Cost of returns	Break-even cut
Outerwear & Coats	1,132	\$155	\$115	\$36,836	1.5 pts (5% of rate)
Jeans	1,646	\$99	\$70	\$32,698	1.7 pts (6% of rate)
Sweaters	1,443	\$74	\$57	\$23,263	2.4 pts (9% of rate)
Fashion Hoodies & Sweatshirts	1,609	\$53	\$42	\$19,233	3.0 pts (10% of rate)
Suits & Sport Coats	681	\$126	\$98	\$18,974	3.0 pts (11% of rate)
Swim	1,486	\$56	\$44	\$18,701	3.0 pts (11% of rate)
Sleep & Lounge	1,430	\$49	\$40	\$16,375	3.5 pts (12% of rate)

Category	Items a year	Average price	Loss per return	Cost of returns	Break-even cut
Shorts	1,461	\$46	\$38	\$15,777	3.6 pts (13% of rate)
Tops & Tees	1,545	\$42	\$34	\$14,829	3.8 pts (13% of rate)
Accessories	1,278	\$44	\$40	\$14,350	4.0 pts (14% of rate)
Whole shop	23,668	\$60	\$47	\$321,593	0.18 pts (0.6% of rate)

Break-even cut = fix cost ÷ (items a year × loss per return). In points of return rate, and as a share of the current rate. The whole shop needs a far smaller cut because the fix is paid once for many more items.

Product review list, top 12

Ranked by expected cost of returns a year, central costs. The full list is in the dashboard downloads.

Product	Items a year	Returned	Price	Cost of returns a year
The North Face Nuptse 2 Jacket Deep Water Blue Mens Outerwear & Coats · The North Face	4	1	\$903	\$743
Robert Graham Men's Barbican Velvet Blazer Suits & Sport Coats · Robert Graham	3	0	\$698	\$420
Arc'teryx Women's Caliber Cardigan Outerwear & Coats · Arc'teryx	3	0	\$699	\$398
Woolrich Arctic Parka DF Outerwear & Coats · Woolrich	2	1	\$990	\$377
The North Face Freedom Mens Ski Pants 2013 Outerwear & Coats · The North Face	2	0	\$903	\$371
The North Face Women's S-XL Oso Jacket Outerwear & Coats · The North Face	2	1	\$903	\$367
The North Face Denali Down Mens Jacket 2013 Outerwear & Coats · The North Face	2	0	\$903	\$344
Canada Goose Men's The Chateau Jacket Active · Canada Goose	2	0	\$815	\$333
Canada Goose Men's The Chateau Jacket Outerwear & Coats · Canada Goose	2	0	\$815	\$317
Sandals Cay Men's Three-Button Cashmere Blend Sport Coat in Gray Navy or Tan Suits & Sport Coats · Sandals Cay	3	1	\$500	\$303
Arc'teryx Men's Sabre Jacket Outerwear & Coats · Arc'teryx	3	0	\$525	\$303
Nike Jordan Retro 11 Bred Bootie Socks Socks · Jordan	2	2	\$903	\$296

Every product gets the shop-wide return rate, because no product has enough sales to show its own. The list ranks by price and volume and is a starting point for checking product pages, not a list of problem products.

Test size guidance shop-wide, randomised by product

A fix such as clearer size guidance is a change to the product page, so the test gives the new guide to half the products, picked at random, and compares return rates. Testing in Outerwear & Coats alone would be the natural start, but at about 22 delivered items a week it can't show a realistic cut in reasonable time. Shop-wide, about 455 items a week, it can.

Test brief: clearer size guidance	
Part	Plan
Change	Clearer size guidance on the product page: a fit note, measurements in centimetres and inches, and how the item fits compared with a common reference.
Who is randomised	Products, half get the new guide, chosen at random. Every product page in the shop is eligible.
Main measure	share of delivered items returned within the return window
Guardrails	sales per product page view; cancellations; complaints or contact about size
Smallest cut worth finding	10% of the current rate (28.4% to 25.6%), well above the break-even cut
When to stop	Fix the length before starting. Don't stop early on a good-looking result.

Weeks needed to see a cut in returns			
80% power, 5% two-sided. Products are randomised, so the count allows for items of the same product being alike.			
Cut in return rate	Points	Outerwear & Coats alone	Whole shop
10%	2.8	387 weeks	17 weeks
20%	5.7	86 weeks	5 weeks
30%	8.5	37 weeks	2 weeks
We assumed items of the same product are a little alike (intra-class correlation 0.01). Fix the length before starting. Don't stop early on a good-looking result.			

The plan was locked before any return rate was compared

The plan (analysis-plan.md) was written and locked after the structural checks and before any return rate was compared between groups. The model choice and every file it depends on were fingerprinted (9568fed61070...) on 2026-10-11 at 20:55 UTC, before the test period was opened. One results file feeds every number on this page.

1 Check the data statuses, dates, prices, costs	2 Build the money bridge, three cost cases, formula check	3 Compare groups rates, ranges, shrinkage, chance	4 Predict, then test frozen choice, scored once on the test period
--	--	--	--

The models <table> <tr> <td>Whole-shop rate</td><td>The training return rate for every item</td></tr> <tr> <td>Category rate</td><td>Each category's rate, pulled toward the shop rate by beta-binomial shrinkage. The baseline to beat.</td></tr> <tr> <td>Logistic regression</td><td>L2, strength picked on validation (C = 0.001), Platt calibration</td></tr> <tr> <td>LightGBM</td><td>31 leaves, learning rate 0.03, 27 rounds (early stopping), Platt calibration</td></tr> </table>	Whole-shop rate	The training return rate for every item	Category rate	Each category's rate, pulled toward the shop rate by beta-binomial shrinkage. The baseline to beat.	Logistic regression	L2, strength picked on validation (C = 0.001), Platt calibration	LightGBM	31 leaves, learning rate 0.03, 27 rounds (early stopping), Platt calibration	Checks we ran - Item records reconcile to their orders - Statuses checked against order age before anything else - Only delivered items past the return window are counted - Customers' earlier returns counted only once settled, before each order - Splits by order date, every item of an order on one side - Ranges by customer bootstrap, 1,000 resamples - Test period opened once, after the fingerprinted freeze - Contribution formula reconciles to the bridge to the cent - verify.py recomputes the main figures from the raw extract
Whole-shop rate	The training return rate for every item								
Category rate	Each category's rate, pulled toward the shop rate by beta-binomial shrinkage. The baseline to beat.								
Logistic regression	L2, strength picked on validation (C = 0.001), Platt calibration								
LightGBM	31 leaves, learning rate 0.03, 27 rounds (early stopping), Platt calibration								

Data dictionary

Field	What it is
Delivered item	An item with status Complete or Returned
Returned	Status Returned. Treated as a full refund of the sale price
Return window	3 days after delivery, the longest any return took
Sale price	What the customer paid. Always equal to the list price here
Product cost	What the shop paid for the item
Fulfilment and fees	Assumed cost to pick, pack and ship, plus a payment fee as a share of the sale
Handling	Assumed cost to receive, check and restock a return
Stock recovered	Assumed share of product cost a returned item is worth back in stock
Cost of returns	Contribution with no returns minus contribution with them
Break-even cut	Drop in return rate at which a fix pays for itself in a year

Decisions log

Decision	Why
Count only delivered items past 3 days	Statuses are random by age, so "not returned yet" can't be inferred
Treat a return as a full refund	There is no refund ledger
Drop 1,430 items dated after the extract	They hadn't happened when the data was taken
Three cost cases, all labelled assumptions	Fulfilment, handling and recovery are not in the data
Products need 30 delivered items to be judged alone	Fewer is noise; none reached it
No model used for the ranking	None beat the category rate on the sealed test
Models refitted once before the test was opened	A code fix changed a frozen file; the test had not been scored
Test shop-wide, not in one category	One category is too small to show a cut in reasonable time

What this data can't tell us

- The data is made up by a program. It shows the method, not how real customers behave.
- There is no refund ledger, return reason, restocking record or size data, so we can't say why items come back.
- Fulfilment, fees, handling and recovery are assumptions. A real shop would put its own numbers in.
- No advertising, overhead or customer acquisition cost is included. Contribution is not profit.
- Nothing was tested. The size-guide test is a plan, not a result.

This is a demonstration of method on made-up data. Every money figure is a scenario built on stated assumptions. It is not a promise of any result, and it is not financial, legal or tax advice.

Every job. Done well.

Dunwell Insights · info@dunwellstudio.com

Synthetic demonstration. Not a client result. Data: TheLook ecommerce (synthetic), Google's made-up clothing shop, via BigQuery public datasets. Customers, products, orders and returns are generated by a program. Google does not endorse Dunwell.