RESEARCH PAPER

The Economics of LEGO: A 50-Year Quantitative Investigation of Product Complexity, Theme Premiums, and Pricing Strategy

Akshay Byagathvalli

November 2026

DATA	18,540 set records · 6,524 priced sets (2005-2025) · Brickset US retail prices
METHODS	Hedonic & quantile regression · mixed models · structural breaks · 12 ML models · SHAP · ablations
HEADLINE	Elasticity 0.640 · +4.6% per minifigure · best MAE $8.06

All estimates are observational associations computed from the data described in Section 3. Prices are original US retail prices, not resale values. Release counts describe catalogue composition, not sales.

Contents

Abstract 3

1. Introduction 3

2. Related Literature 4

3. Data 6

4. Methodology 9

5. Exploratory Results 11

6. Econometric Results 29

7. Machine Learning Results 38

8. Robustness Analysis 48

9. Discussion 49

10. Limitations 50

11. Conclusion 51

References 52

Appendix 54

Abstract

How much of the variation in LEGO retail prices can be explained by observable product characteristics, and has that relationship changed over half a century? This study links the Kaggle *LEGO Sets, Themes, and Minifigures (1975-2025)* release (18,540 set records derived from Rebrickable) to original US retail prices from two Brickset-derived archives by exact set identifier, producing 8,495 priced conventional retail sets and a densely priced modern sample of 6,524 sets released 2005-2025. Methods include robust and quantile trend tests, hedonic regressions with heteroskedasticity- and cluster-robust inference, spline and mixed-effects models, propensity-score sensitivity analysis, sup-F structural-break tests, theme diversification indices, and twelve machine-learning models evaluated under group-disjoint random and strictly temporal holdouts with controlled feature ablations. Median conventional set size rose 375% between 1975-1984 and 2016-2025 (bootstrap 95% CI [283%, 432%]), and roughly 63% of the annual growth rate persists within themes. Price scales sub-proportionally with piece count: the log-log elasticity is 0.640 (95% CI 0.624-0.656), so doubling parts is associated with a 56% higher price. Each additional minifigure carries a conditional association of 4.6% (bootstrap CI [4.1%, 5.1%]). Theme identity is the largest source of conditional variation after size, raising R^2 from 0.699 to 0.882; a licensed-theme premium is specification-dependent, near zero without and 10.5% [1.0%, 20.8%] with a control for large-element (Duplo-type) formats. Real price per piece roughly halved between 2005 and 2025. The best held-out model (Extra Trees) predicts prices with MAE $8.06 and R^2 0.916; adding theme reduces linear-model MAE by 32%. All results are observational associations, not causal effects of product attributes on price.

Keywords: hedonic pricing, product differentiation, LEGO, machine learning, structural change, quantile regression, SHAP

1. Introduction

LEGO sets are an unusually well-documented consumer product system. Each set has a public identifier, a release year, a theme, a part inventory and, for most sets sold in the United States since the mid-2000s, a published retail price. The catalogue spans small polybags and toddler sets through collector models with more than ten thousand parts, and themes ranging from LEGO-owned lines such as City and Technic to licensed franchises such as Star Wars and Harry Potter. Because the core input of every set is the same plastic element system, LEGO offers a natural laboratory for asking how a manufacturer prices variation in size, figure content and intellectual property within one product family.

Hedonic pricing theory treats a differentiated product as a bundle of characteristics whose implicit prices can be recovered from cross-sectional variation in market prices (Court, 1939; Griliches, 1961; Lancaster, 1966; Rosen, 1974). In the LEGO case the obvious characteristics are piece count, minifigure inclusion, theme and release period. Popular discussion often assumes that licensed sets carry a premium and that price per piece is a stable yardstick of value. Neither assumption has been examined with a long, reproducible, set-level panel, and both conflate distinct quantities: price per piece mechanically falls with set size if prices scale less than proportionally, and theme premiums can reflect element format or target age rather than licensing.

This paper answers the central question: **to what extent do piece count, minifigure inclusion, theme identity and release year explain variation in LEGO set prices, and how have their relative contributions changed between 1975 and 2025?** Six secondary questions structure the analysis:

- **RQ1 Complexity evolution.** How have set size and minifigure content evolved, and is growth broad-based or driven by extreme products?
- **RQ2 Theme premiums.** Do themes differ in conditional price after adjusting for size, minifigures and year, and do licensed themes differ from LEGO-owned themes?

- **RQ3 Minifigure economics.** What is the conditional price association of an additional minifigure, and is it nonlinear or theme-dependent?
- **RQ4 Pricing evolution.** Has the price-size relationship changed over time in nominal and real terms, and are apparent changes structural or compositional?
- **RQ5 Predictive modelling.** How accurately can retail price be predicted from observable characteristics, and which information contributes most?
- **RQ6 Portfolio evolution.** How have theme concentration, diversification and longevity in the release catalogue changed?

The contributions are fourfold. First, the study documents a transparent, verifiable price join between the Rebrickable-derived Kaggle release, which contains no price field, and Brickset archives, with explicit treatment of price coverage and selection. Second, it estimates a family of hedonic models with appropriate inference for theme-level regressors, showing where common claims (licensed premiums, falling prices) survive and where they do not. Third, it quantifies the predictive value of theme and minifigure information with controlled ablations rather than relying on feature-importance rankings alone. Fourth, all 43 figures, the tables and this paper are generated by a single reproducible pipeline whose outputs are tested for consistency.

2. Related Literature

Hedonic pricing. Court (1939) and Griliches (1961) introduced regressions of product prices on measurable characteristics to adjust price indexes for quality change in automobiles. Rosen (1974) provided the equilibrium interpretation: hedonic coefficients trace the market's implicit price schedule for characteristics, which mixes supply and demand conditions and need not equal any individual consumer's willingness to pay. Triplett (2006) synthesizes practice in official statistics, including the near-universal use of log-price specifications and the interpretation problems created by correlated characteristics. This study adopts that framework while emphasising that LEGO prices are set by a single firm, so coefficients describe that firm's pricing schedule rather than a competitive market equilibrium.

Product differentiation and brand premiums. Lancaster (1966) modelled goods as bundles of characteristics, which motivates separating the physical content of a set (parts, figures) from its positioning (theme). Licensed products are a familiar form of differentiation, and licensing fees are commonly assumed to be passed through to consumers. Because licence terms are confidential, the present data can only test whether licensed themes are priced differently conditional on observables, not why.

Complexity and product evolution. Business histories describe LEGO's expansion into licensing in the late 1990s, financial difficulties in the early 2000s and subsequent restructuring (Robertson & Breen, 2013). These accounts motivate exploratory period definitions, but this paper does not impose those dates as break points; it searches for breaks statistically and treats narrative history as context only.

LEGO-specific empirical work. Dobrynskaya and Kishilova (2022) study secondary-market returns to retired LEGO sets and find returns that compare favourably with conventional assets; that work concerns resale value, a different quantity from the original retail prices analysed here. Bartneck, Obaid and Zawieska (2013) analyse changes in minifigure facial expressions over time, illustrating the use of the LEGO catalogue as longitudinal data. We are not aware of a peer-reviewed hedonic study of LEGO original retail prices with the scope of the present analysis.

Predictive pricing and interpretability. Tree ensembles such as random forests (Breiman, 2001) and gradient boosting (Friedman, 2001), with modern implementations in XGBoost (Chen & Guestrin, 2016), LightGBM (Ke et al., 2017) and CatBoost (Prokhorenkova et al., 2018), routinely outperform linear models on tabular price prediction. Interpretation relies on SHAP values (Lundberg & Lee, 2017), partial dependence and ICE curves

(Goldstein et al., 2015) and accumulated local effects (Apley & Zhu, 2020). These tools describe a fitted model, not the data-generating process, which is why the present study pairs them with ablations and with interpretable econometric models.

3. Data

3.1 Provenance and schema

The primary dataset is the Kaggle release *LEGO Sets, Themes, and Minifigures (1975-2025)* (Hacinacar, 2025), which derives from the Rebrickable catalogue. It contains three tables: **sets** (18,540 rows; set number, name, year, theme, parent theme, part count, minifigure count, distinct colours, image URL), **themes** (464 rows) and **minifigures** (15,790 rows; identifier, name, part count and the number of sets in which each figure appears). Programmatic schema discovery confirmed that `set_num` is a unique key, that every set matches a (theme, parent theme) pair in the theme table, that there are no missing cells, no zero-part sets and no years outside 1975-2025, and that `has_minifigs` is exactly `num_minifigs > 0`. Two schema facts shaped the design. First, **the release contains no price field**. Second, it contains **no set-to-minifigure edge list**, only per-set counts and per-figure appearance totals.

3.2 Supplementary price data

Original US retail prices were obtained from two public, Brickset-derived archives and joined by exact set identifier: the `legosets` data file of the *brickset* R package (Bryer, 2026; updated January 2026 to include 2025 sets) and the 2015 Brickset snapshot distributed with the *lego* R package (Kross, 2015). The 2026 archive matched 16,181 Kaggle sets and supplied 6,842 prices; the 2015 snapshot filled 2,776 additional, mostly pre-2005, prices. Where both sources report a price (2,496 sets), they agree exactly in 98.9% of cases, and the 2026 archive is given priority. 133 zero prices (free promotional items) were recoded as missing rather than treated as prices. Brickset release years agree with Rebrickable years for 96.9% of matched sets; piece counts agree exactly for 70.0% because the two catalogues count spare and figure parts differently (median absolute difference 0%). The price variable is the **original US retail price**, not resale or current market value. Prices were deflated to 2025 dollars with the annual mean of the seasonally adjusted CPI-U (FRED series CPIAUCSL; U.S. Bureau of Labor Statistics, 2026), so that real price = nominal price × CPI(2025)/CPI(t), with CPI(2025) = 321.96.

3.3 Cleaning and analytical samples

Extreme sets were not removed for being unusual; very large sets are genuine products. Instead, nonstandard merchandise was identified with documented rules: parent themes that are not construction sets (gear, books, service and bulk packs, promotional, brand-store, education, Legoland, FIRST LEGO League, collectible minifigures), Brickset categories other than normal sets (gear, books, collections, extended parts packs), names indicating promotional items or bundles, and electronic components (Power Functions; Mindstorms items with at most ten parts) whose one-part counts would otherwise produce price-per-piece values above $100. Five nested samples follow (Table 1).

Table 1. Analytical samples and representation by decade.

Sample	N	First	Last	Themes	Median parts	1970s	1980s	1990s	2000s	2010s	2020s
full	18,540	1975	2025	144	59	399	1214	2104	3768	6260	4795
conventional	12,211	1975	2025	131	116	269	793	1619	2733	3813	2984
priced	8,495	1976	2025	125	191	8	192	770	1936	3190	2399
modern	6,524	2005	2025	103	231	0	0	0	935	3190	2399
theme_comparable	5,974	2005	2025	41	223	0	0	0	799	2986	2189

Table 2. Sample construction and exclusion log.

Step	Rule	Removed	Remaining
full_historical	All 18,540 Kaggle set records 1975-2025	0	18,540
conventional_retail	Drop nonstandard parent themes (gear, books, service/bulk packs, promotional, brand-store, education, Legoland, FLL, collectible minifigures)	4,537	14,003
conventional_retail	Drop Brickset categories Gear/Book/Collection/Extended/Random/Other	1,522	12,481
conventional_retail	Drop names indicating promotional items, key chains, magnets, bundles, kits, collections	191	12,290
conventional_retail	Drop electronic components (Power Functions; Mindstorms items with <= 10 parts)	79	12,211
price_observed	Conventional retail sets with an observed original US retail price > 0	3,716	8,495
modern_era	Price-observed sets released 2005-2025 (dense price coverage)	1,971	6,524
theme_comparable	Modern priced sets in parent themes with >= 30 priced sets	550	5,974

Price observation is strongly selective (Figure 1). Among conventional sets the observed share rises from 3.0% in the 1970s to 83.7% in the 2010s ($\chi^2 = 2229$, $p < 0.001$), differs by theme ($\chi^2 = 1576$, $p < 0.001$) and is higher for larger sets (median 191 versus 40 parts; Mann-Whitney $p < 0.001$). Within 2005-2025, 81.1% of conventional sets are priced, and a logit with year effects shows that observation still rises with log parts ($\beta = 0.633$, $p < 0.001$) and minifigures ($\beta = 0.375$). Pricing analyses therefore focus on 2005-2025, are labelled as conditional on price observation, and are re-estimated with inverse-probability weights in Section 8.

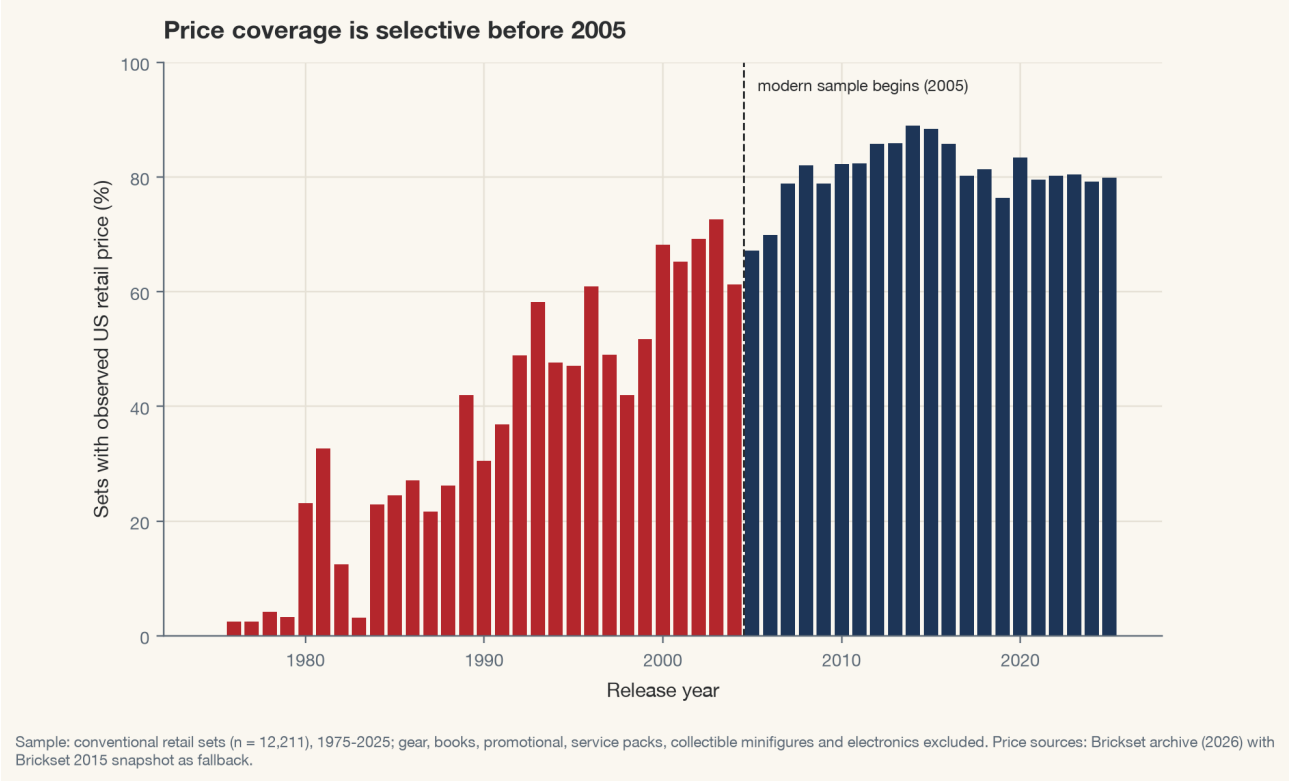


Figure 1. Share of conventional sets with an observed original US retail price, by release year. Coverage is sparse and selective before 2005, which motivates the modern pricing sample.

3.4 Descriptive statistics

Table 3. Descriptive statistics for key variables.

Variable	N	Mean	Median	SD	IQR	Skew
Parts (conventional, 1975-2025)	12,211	301.5	116	552.2	307	6.10
Minifigures (conventional)	12,211	1.54	1	2.36	2	3.10
Price, nominal USD (2005-2025)	6,524	47.41	29.99	62.57	41.25	4.64
Price, 2025 USD (2005-2025)	6,524	58.78	35.77	74.46	52.68	4.43
Price per part, USD (2005-2025)	6,524	0.215	0.111	0.574	0.066	16.66
Parts (2005-2025 priced)	6,524	444.8	231	692.6	445	5.10

All size and price variables are strongly right-skewed, so models use logarithms. In the modern sample, Spearman correlations with log price are 0.86 for log parts, 0.44 for minifigures, 0.66 for distinct colours and 0.18 for the licensed indicator; parts and minifigures are themselves correlated (0.46), which is why unadjusted minifigure comparisons overstate their independent association.

4. Methodology

4.1 Hypotheses

Five pre-specified null hypotheses were tested: **H1** complexity has no systematic relationship with release year; **H2** conditional prices are equal across themes; **H3** minifigure count has no conditional association with price; **H4** the relationship between characteristics and price is stable over time; **H5** licensed and LEGO-owned themes have equal conditional expected prices. Every test reports a statistic, p-value, effect size, 95% confidence interval and sample size, and families of theme contrasts are corrected with the Holm procedure (Benjamini-Hochberg q-values are also saved).

4.2 Econometric specifications

Let *i* index sets, *k(i)* the parent theme and *t(i)* the release year. The baseline, log-log and hedonic models are

$$P_i = \beta_0 + \beta_1 \mathrm{Parts}_i + \varepsilon_i \tag{1}$$

$$\log P_i = \beta_0 + \beta_1 \log \mathrm{Parts}_i + \varepsilon_i \tag{2}$$

$$\log P_i = \beta_0 + \beta_1 \log \mathrm{Parts}_i + \beta_2 \mathrm{Minifigs}_i + \beta_3 (t_i - 2015) + \beta_4 \mathrm{Licensed}_{k(i)} + \varepsilon_i \tag{3}$$

In the log-log model β_1 is an elasticity; $\beta_1 = 1$ would mean proportional pricing, so that price per piece is invariant to size. The theme fixed-effects model replaces the licensed indicator and linear trend with theme and year effects:

$$\log P_i = \beta_1 \log \mathrm{Parts}_i + \beta_2 \mathrm{Minifigs}_i + \gamma_{k(i)} + \delta_{t(i)} + \varepsilon_i \tag{4}$$

Because licensing is defined at the parent-theme level, it is perfectly collinear with γ and is not estimated in this model. Theme contrasts γ are expressed relative to City, a large LEGO-owned theme present throughout the modern period, and converted to percentages as 100·(exp(γ) − 1); log coefficients are never read directly as percentages. The nonlinear model replaces linear terms with cubic B-splines (5, 4 and 5 degrees of freedom for log parts, minifigures top-coded at 15, and year); the interaction model adds log parts × era (2005-09, 2010-14, 2015-19, 2020-25) and minifigures × licensed; quantile regressions (Koenker & Bassett, 1978) estimate the hedonic model at the 10th, 25th, 50th, 75th and 90th conditional percentiles; and a linear mixed model with theme random intercepts and random log-parts slopes provides partial pooling:

$$\log P_i = \mathbf{x}_i' \boldsymbol{\beta} + u_{0k(i)} + u_{1k(i)} \log \mathrm{Parts}_i + \varepsilon_i, \quad (u_{0k}, u_{1k}) \sim N(\mathbf{0}, \Sigma) \tag{5}$$

Inference uses HC3 heteroskedasticity-robust standard errors (MacKinnon & White, 1985). Because the licensed indicator varies only across themes, its standard errors are clustered by parent theme (Cameron & Miller, 2015) and complemented by a theme-cluster bootstrap; theme fixed effects themselves use HC3, since clustering on the same dimension as the dummy is uninformative. Variance inflation factors, Breusch-Pagan tests, Q-Q plots and Cook's distances were inspected, and the hedonic model was re-estimated without observations with Cook's D > 4/N. Licensed-theme comparisons are additionally estimated by inverse-probability weighting and 1:1 nearest-neighbour propensity matching on log parts, minifigures, year and element format (Rosenbaum & Rubin, 1983).

4.3 Trend, structural-change and diversification methods

Complexity trends are tested with Spearman and Kendall correlations, the Mann-Kendall test with Sen slope on annual medians and 90th percentiles, Huber robust regression of log parts on year, quantile regression, pairs-bootstrap intervals and a theme fixed-effects regression that isolates within-theme growth. For the pricing relationship, a single unknown break in intercept and elasticity is tested with the sup-F statistic over candidate years with 15% trimming (Andrews, 1993); because the statistic is a maximum over many candidate dates, its p-value comes from a 299-replication fixed-regressor bootstrap under the no-break null (Hansen, 2000). Annual series are segmented with PELT (Killick et al., 2012) using a minimum segment of five years and a BIC-type penalty. Release concentration in year *t* is measured by

$$\mathrm{HHI}_t = \sum_k s_{kt}^2, \qquad H_t = -\sum_k s_{kt} \log s_{kt} \tag{6}$$

where s is theme k's share of sets released in year t (Rhoades, 1993; Shannon, 1948), together with effective numbers 1/HHI and exp(H), entry and exit rates and Kaplan-Meier theme survival (Kaplan & Meier, 1958). These indices describe catalogue composition, not sales.

4.4 Machine-learning framework

The prediction target is log retail price; predictions are back-transformed to dollars with Duan's (1983) smearing factor estimated from five-fold out-of-fold residuals within the training data, and accuracy is reported in dollars (MAE, RMSE, R^2, median absolute error, WAPE, RMSLE) and on the log scale. Candidate features are log parts, minifigures, year, distinct colours, minimum age (with a missing indicator), the licensed indicator, parent theme and packaging. Price per piece and any other price-derived variable are excluded by an automated leakage test. Twelve models are compared: a median dummy, OLS, ridge, lasso, elastic net, random forest, extra trees, histogram gradient boosting, XGBoost, LightGBM, CatBoost and a generalised additive model (Hastie & Tibshirani, 1990). Two designs are used. The **random holdout** draws 20% of sets as a test set with all variants of a set number kept together (group split; 1,305 test sets). The **temporal holdout** trains on 2005-2019 (4,125 sets), tunes on 2020-2021 (765), and evaluates once on 2022-2025 (1,634). Hyperparameters are tuned with Optuna's TPE sampler (Akiba et al., 2019) using three-fold group cross-validation within the training set (random design) or the validation years (temporal design); preprocessing is fitted inside each training fold. An expanding-window rolling-origin evaluation (test years 2012-2025) and feature ablations A-G complete the design. Ablations hold model, hyperparameters and splits fixed and are repeated over five group splits.

5. Exploratory Results

5.1 Catalogue growth and portfolio composition (RQ6)

Annual releases grew from 40 records in 1975 to 921 in 2025, the largest annual total in the data (Figure 2, Figure 3). Conventional retail sets rose from 24 to 548. The composition of the catalogue changed more than its size (Figure 4, Figure 5): Town and Duplo dominated the 1980s, Star Wars appears from 1999, City replaces Town from the mid-2000s, and Friends and Ninjago become major lines in the 2010s.

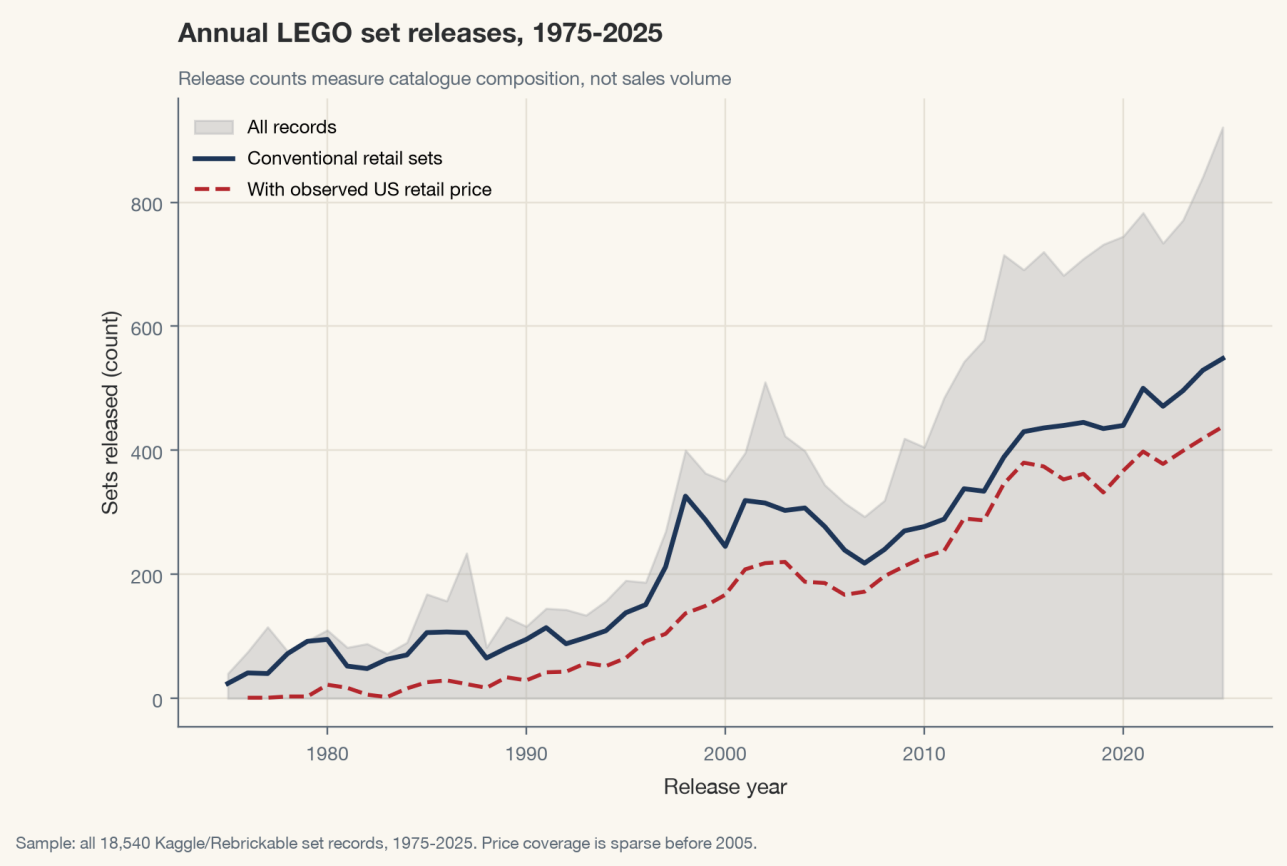


Figure 2. Annual releases for all records, conventional retail sets and price-observed sets.

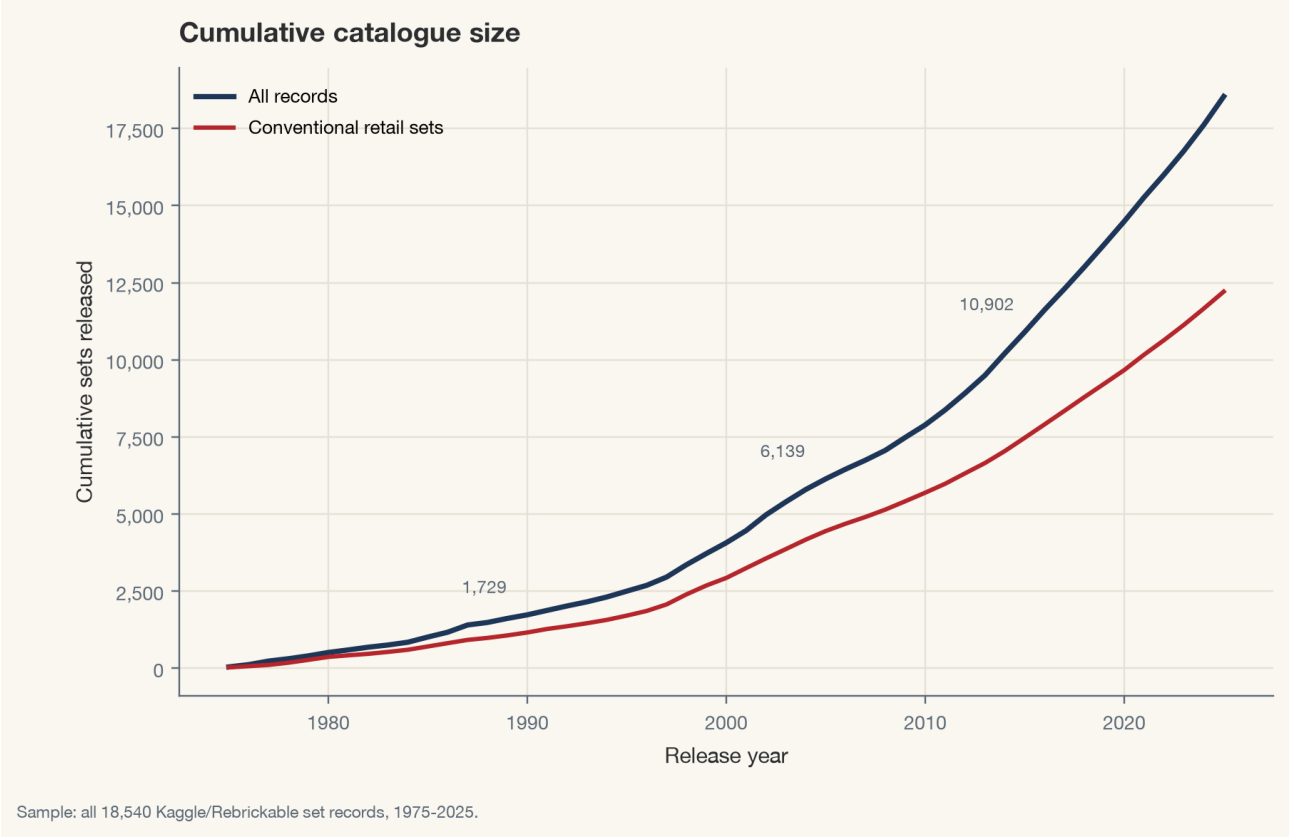


Figure 3. Cumulative catalogue size.

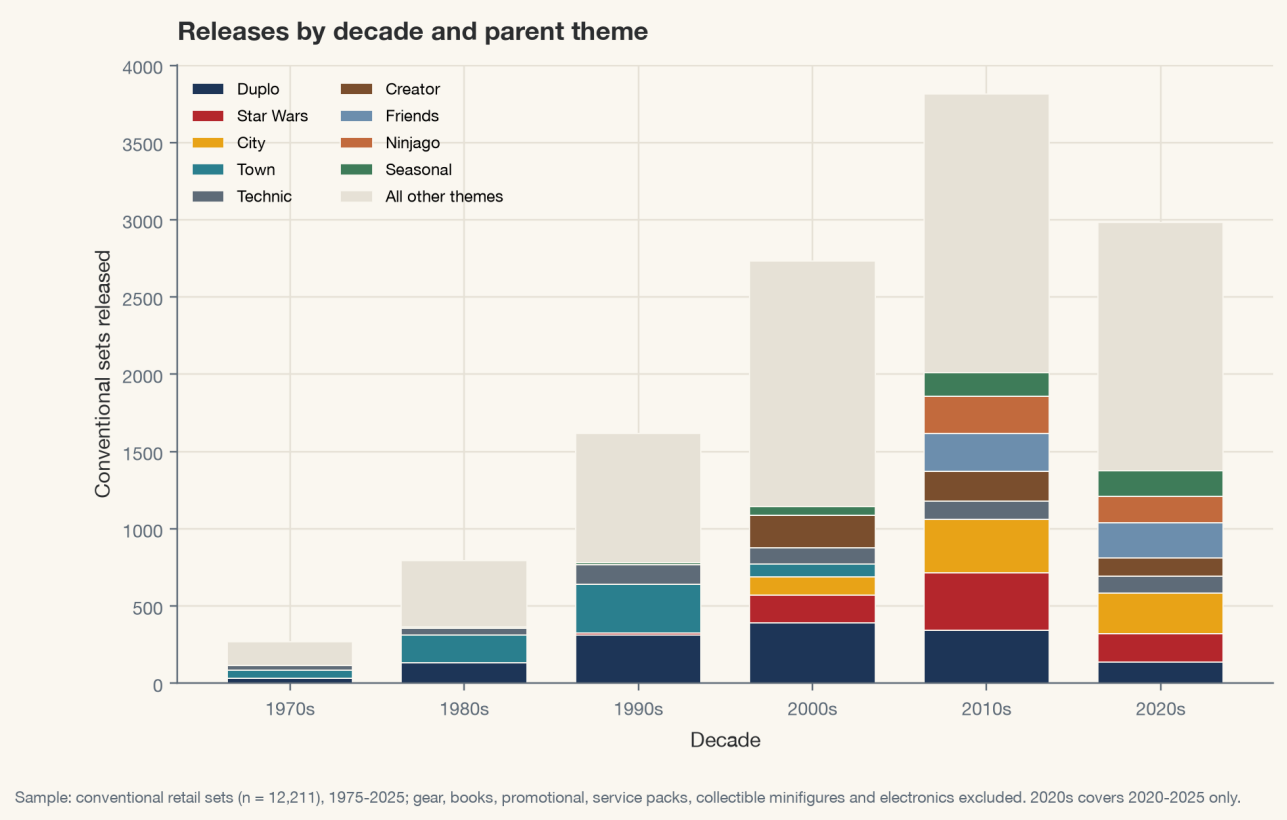


Figure 4. Conventional releases by decade and parent theme.

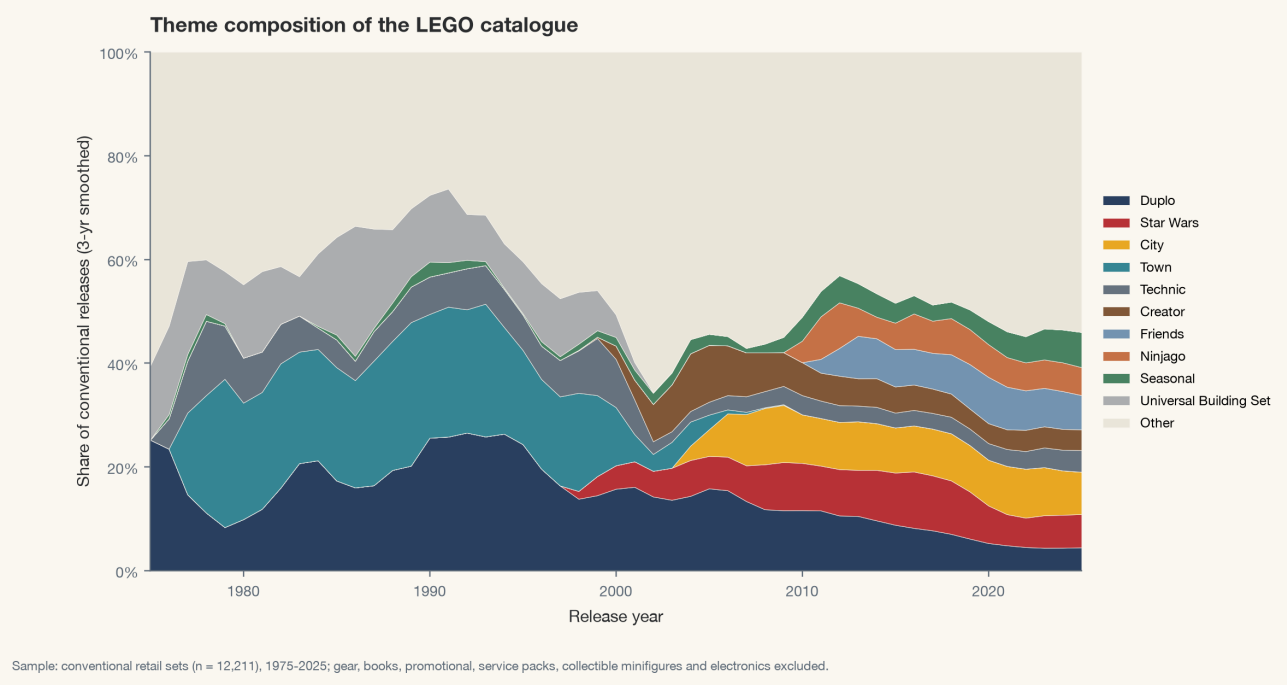


Figure 5. Theme shares of conventional releases (three-year smoothing).

Release concentration fell sharply (Figure 6). The theme HHI averaged 0.211 in 1975-1979 and 0.047 in 2021-2025, while the entropy-based effective number of themes rose from 5.9 to 25.4 and the number of active parent themes from 8.0 to 34.8 (peak 39 in 2025). Themes enter and exit continually: the mean annual entry rate is

10.3% and the exit rate 7.6% of active themes. Kaplan-Meier estimates put median theme lifespan at 6 years, with 43% of themes still active ten years after launch (Figure 7, Figure 8). Diversification and median set size are positively correlated in levels (ρ = 0.65) but not in first differences (ρ = -0.10, p = 0.507), so the association reflects shared long-run trends rather than year-to-year co-movement.

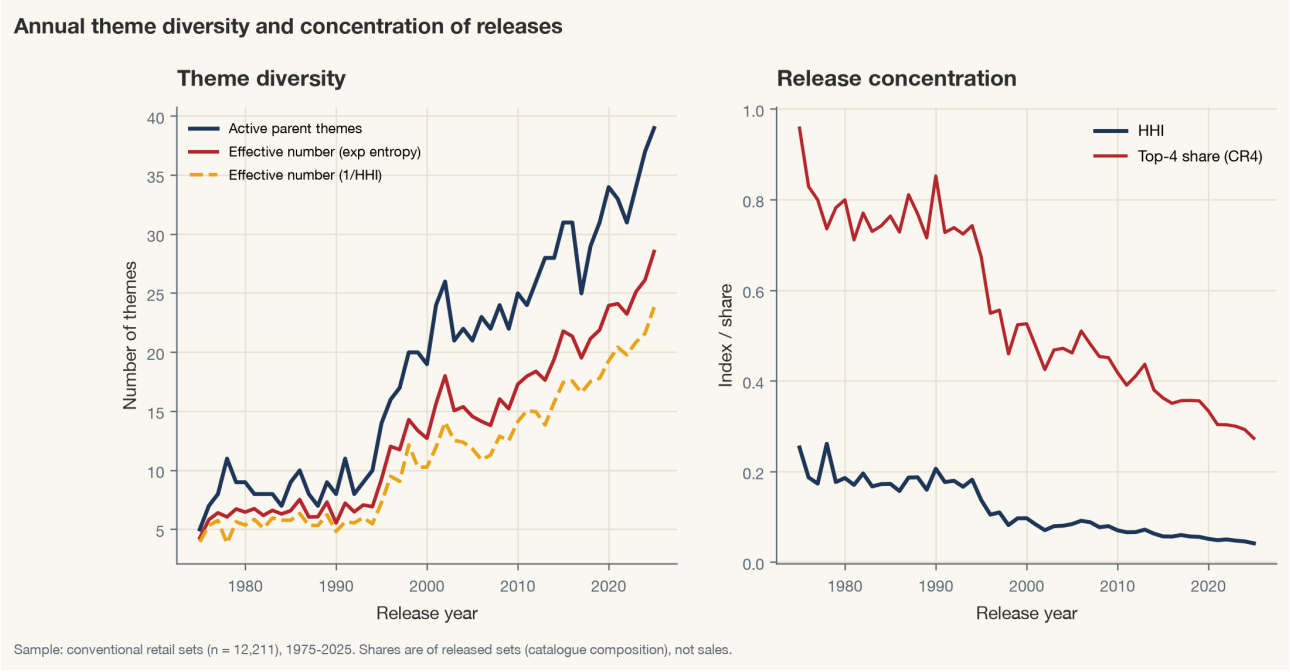


Figure 6. Active themes, effective numbers of themes and release concentration (HHI, CR4).

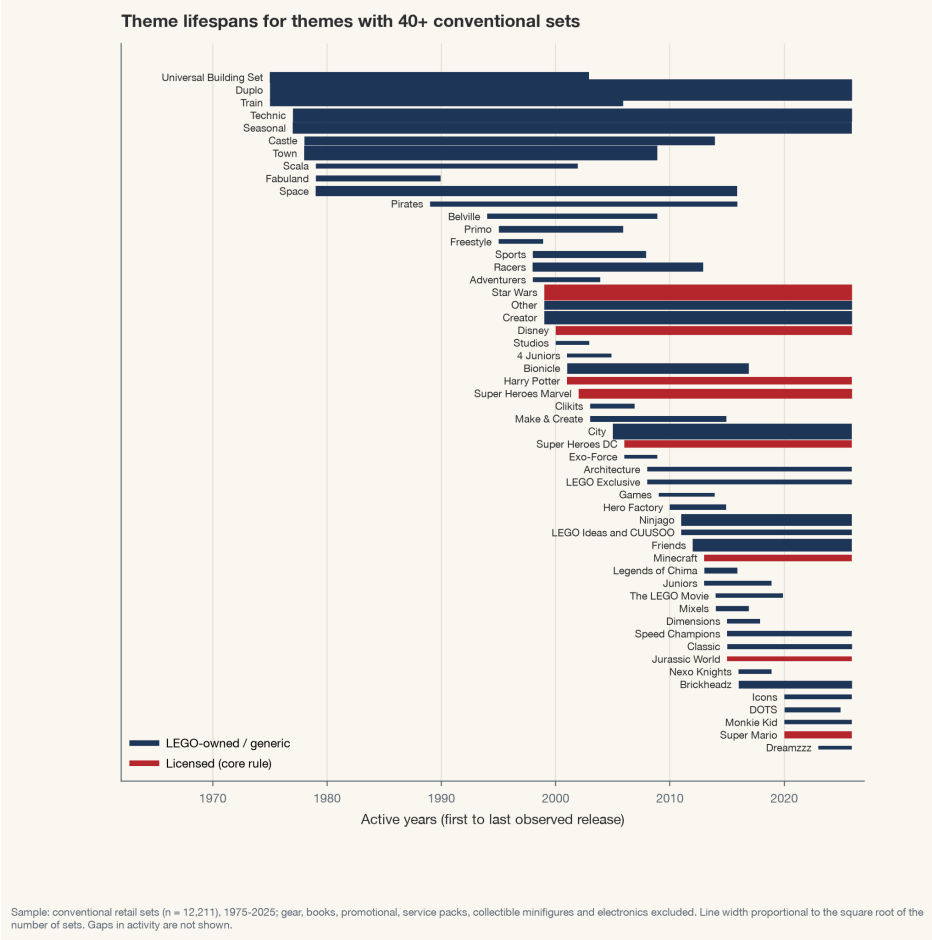


Figure 7. Active periods of parent themes with at least 40 conventional sets.

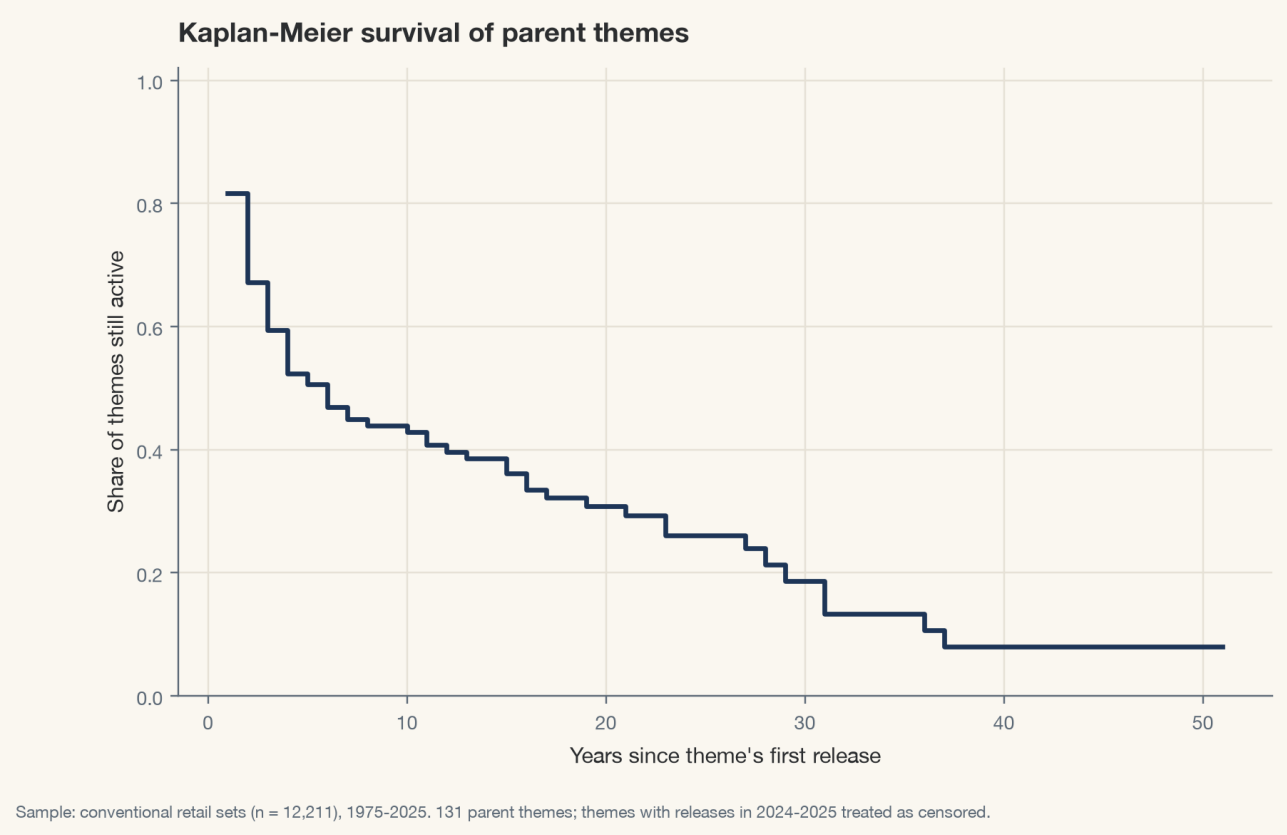


Figure 8. Kaplan-Meier survival of parent themes.

5.2 Complexity (RQ1)

Set size has grown substantially and across the distribution (Figure 9, Figure 10, Figure 11). The median conventional set had 48 parts in 1975-1984 and 228 in 2016-2025, a change of 375% (bootstrap 95% CI [283%, 432%]). The mean rose 301% (from 117 to 468) and the 90th percentile 217% (from 338 to 1072). Sets with at least 1,000 parts were absent in 1975-1984 and made up 11.5% of releases in 2016-2025 (Figure 12). By decade, median size was 1970s: 68, 1980s: 48, 1990s: 49, 2000s: 63, 2010s: 146, 2020s: 262.

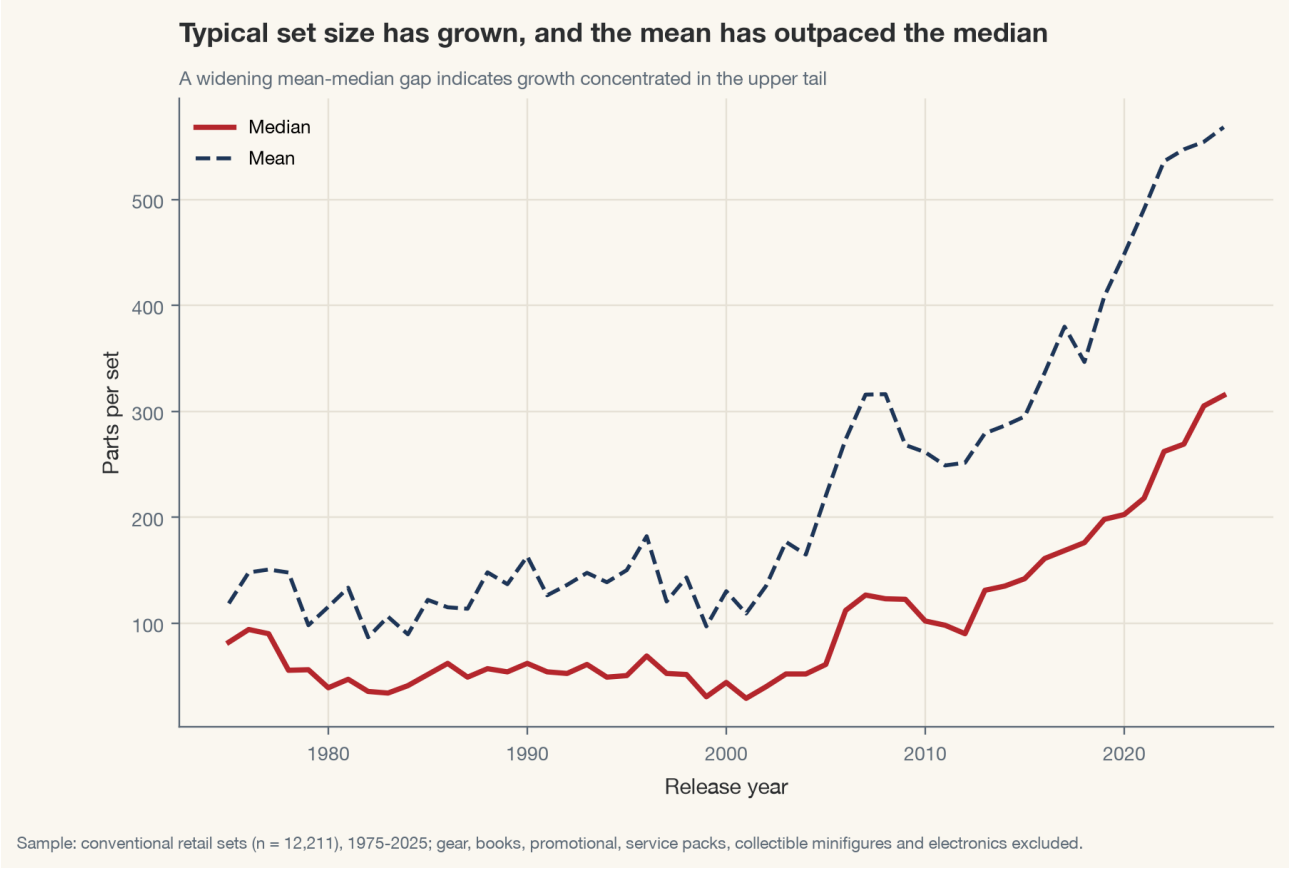


Figure 9. Annual median and mean parts per conventional set.

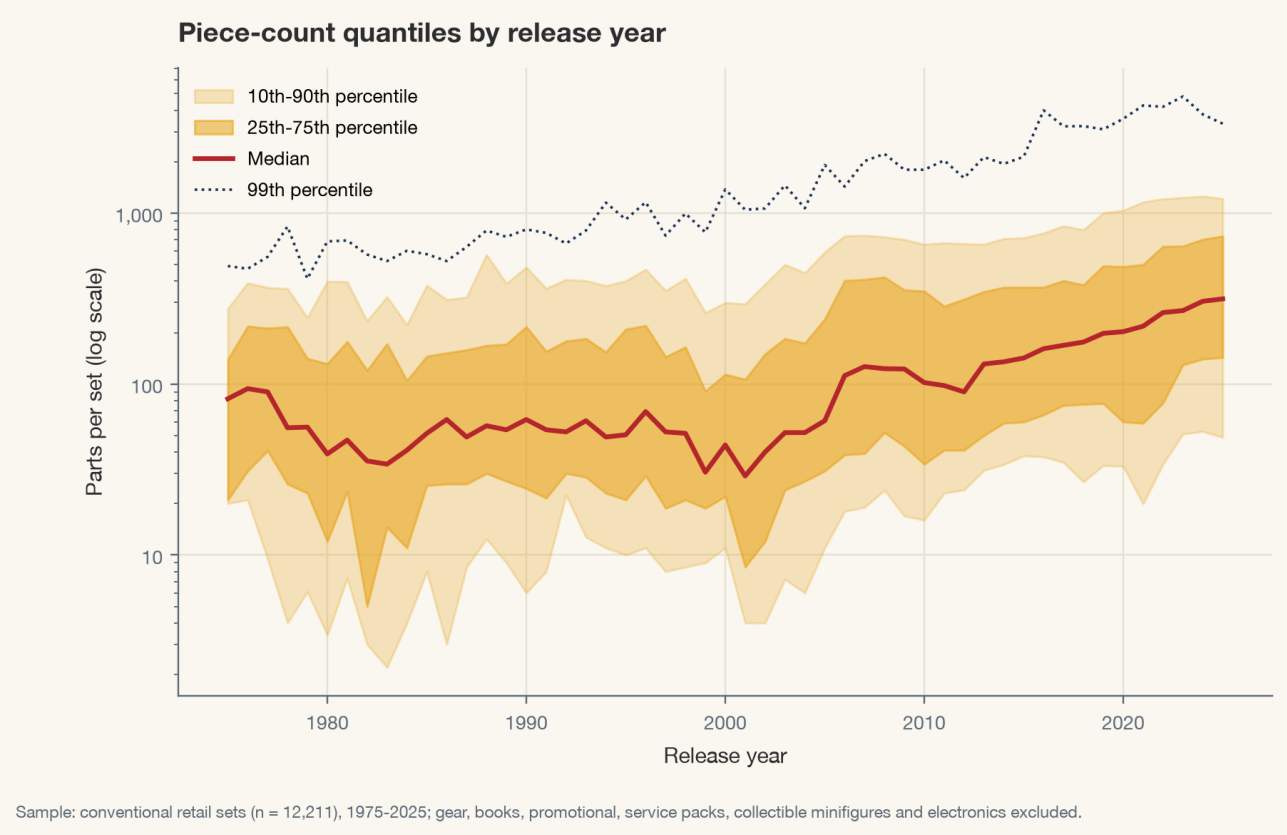


Figure 10. Annual piece-count quantiles (log scale).

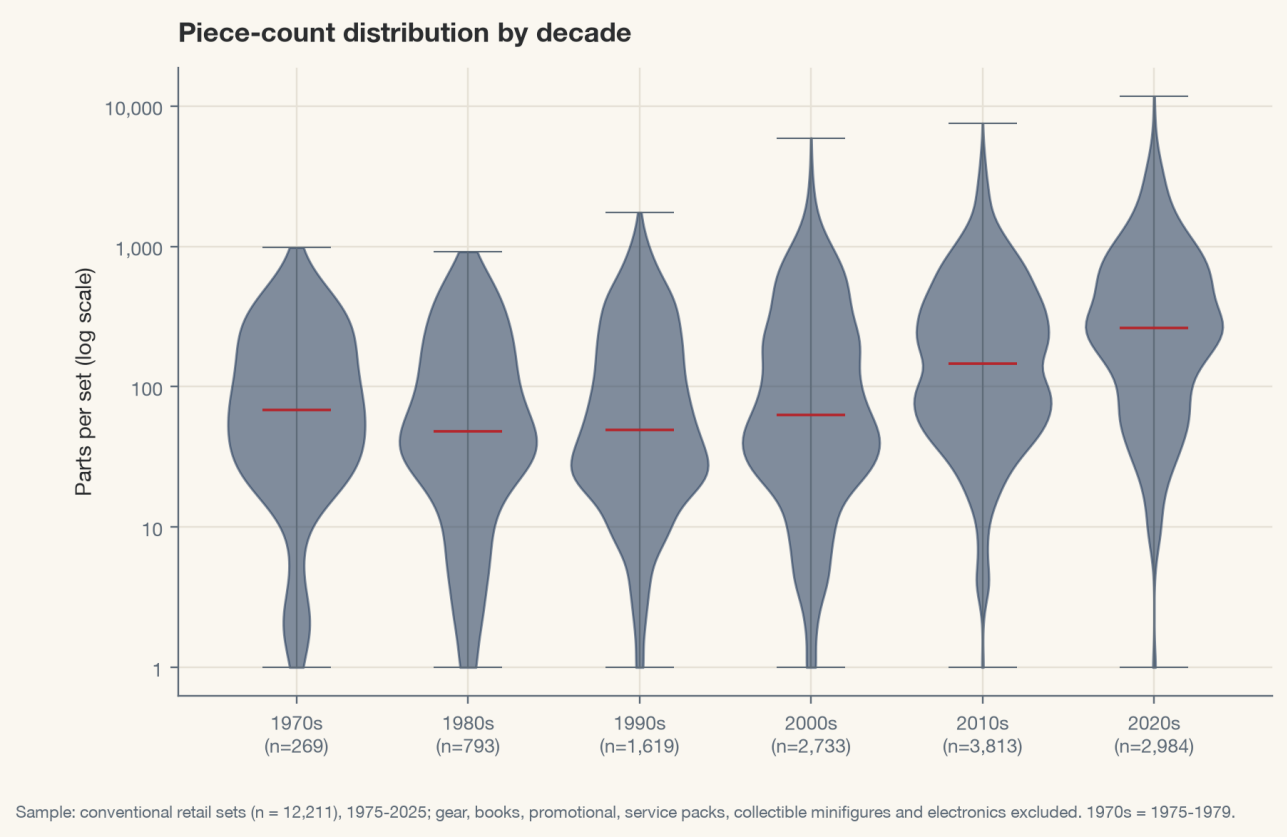


Figure 11. Piece-count distribution by decade.

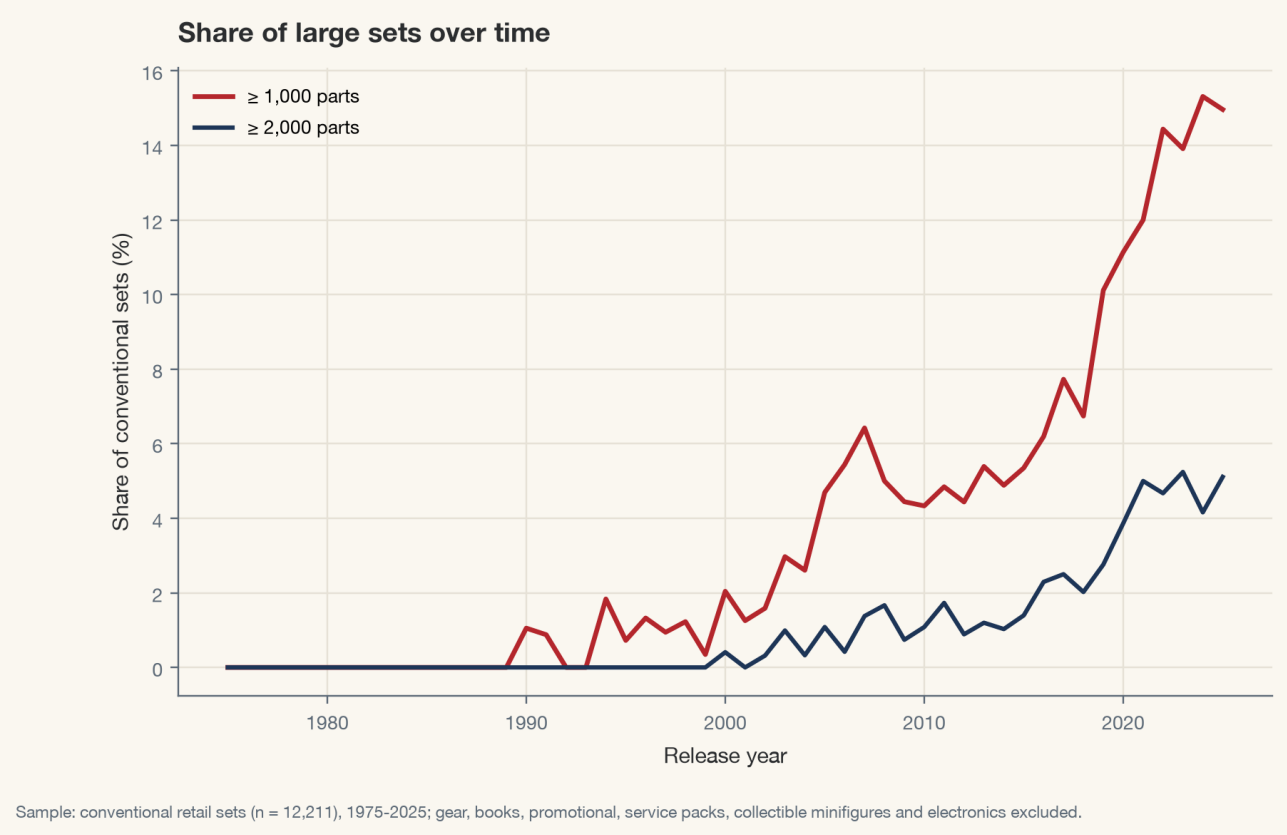


Figure 12. Share of conventional sets with at least 1,000 and 2,000 parts.

Growth is broad-based. Of the 11 parent themes with at least 20 sets both before and after 2000, 9 have a larger median after 2000. The theme heatmap (Figure 13) shows growth within long-running lines rather than only through the arrival of new large-format themes: median Technic sets grew from 179 to 583 parts, Star Wars from 158 to 262, and Town from 66 to 162, while the exceptions (Primo and Sports) are small or short-lived. Minifigure inclusion also rose: 40% of 1975-1984 sets contained a figure versus 60% in 2016-2025 (Figure 14).

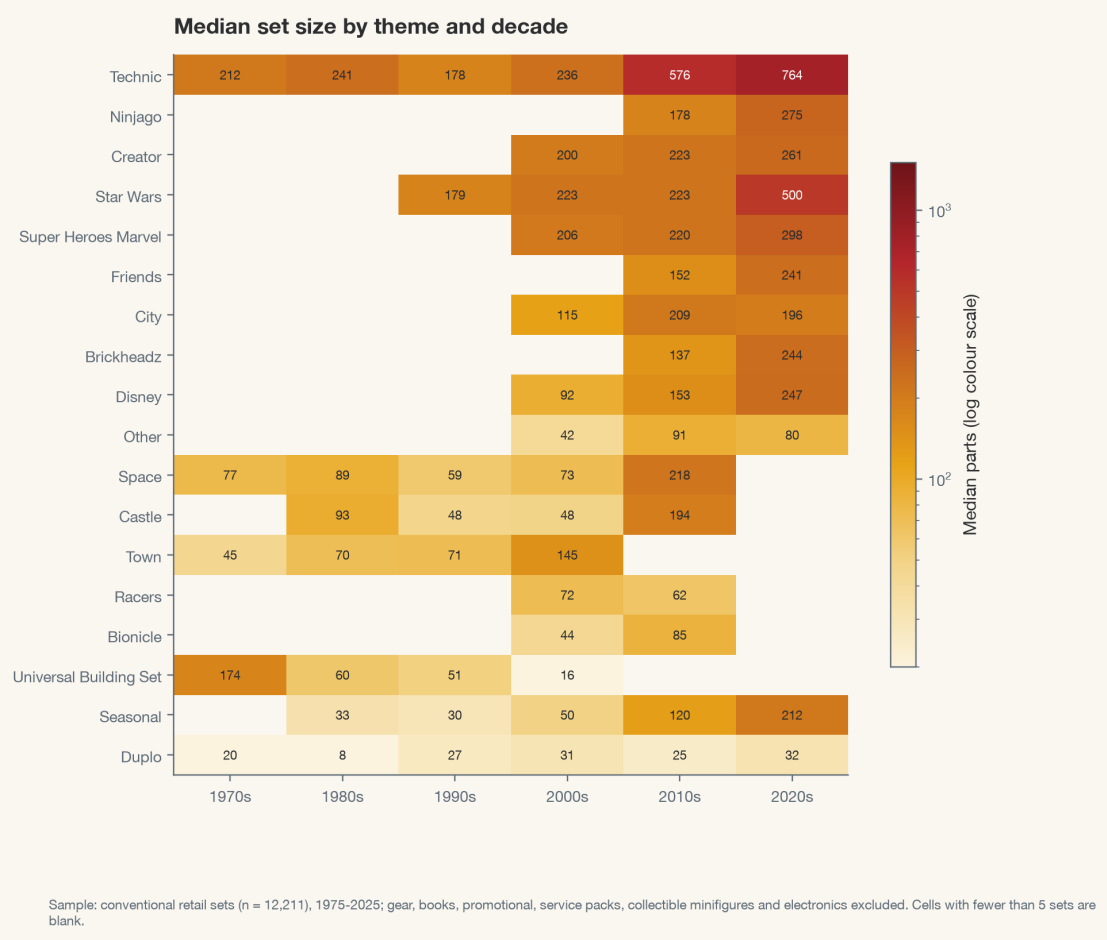
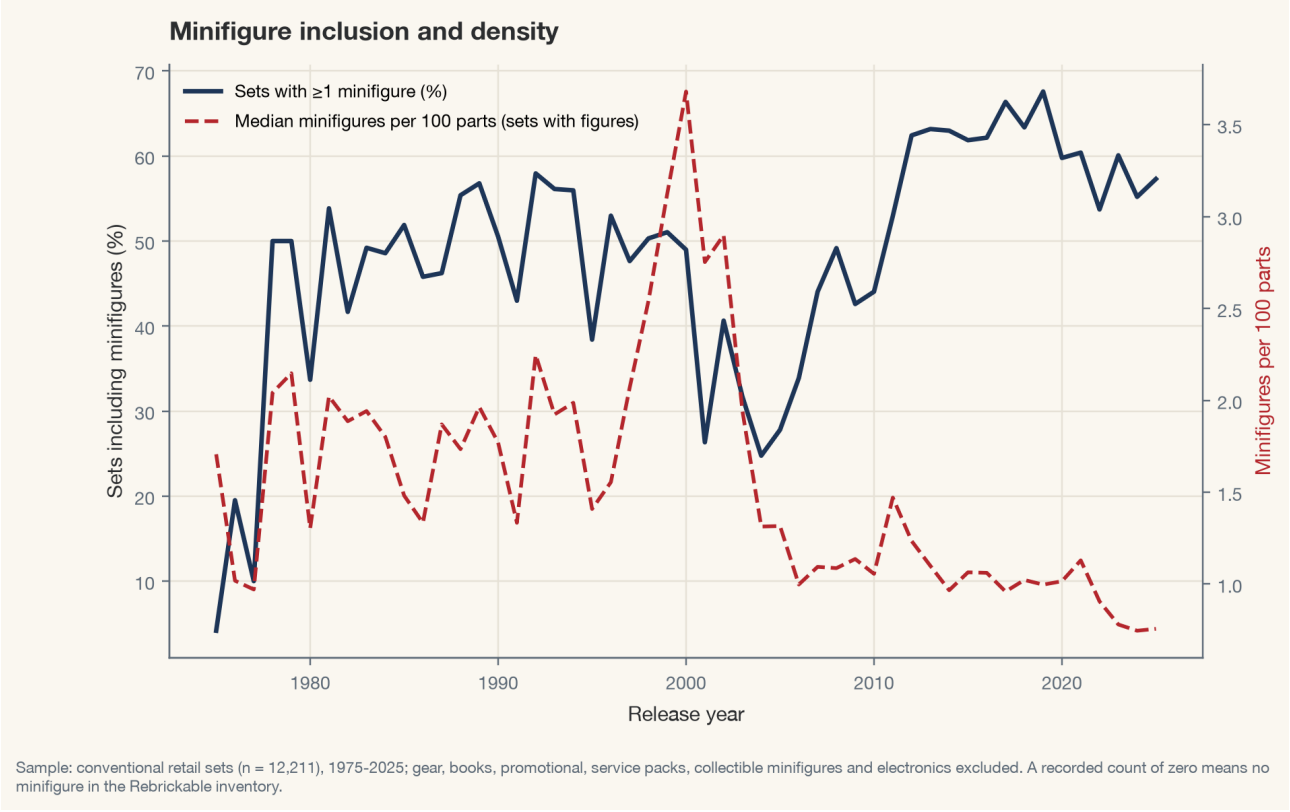


Figure 13. Median parts by theme and decade (cells with fewer than five sets blank).



Sample: conventional retail sets (n = 12,211), 1975-2025; gear, books, promotional, service packs, collectible minifigures and electronics excluded. A recorded count of zero means no minifigure in the Rebrickable inventory.

Figure 14. Share of sets with minifigures and minifigure density.

5.3 Prices (RQ4)

In the modern sample, the median nominal price is $29.99 and the median price per part $0.111. Price rises steeply but sub-proportionally with size (Figure 15, Figure 16): median prices by size class are <100 parts $9.99 (n = 1,898), 100-299 parts $19.99 (n = 1,914), 300-999 parts $49.99 (n = 2,046), 1000+ parts $139.99 (n = 666) (Figure 17), while median price per part falls from $0.192 to $0.092 across the same classes. The latter pattern is a direct consequence of an elasticity below one and means raw price-per-piece comparisons across themes of different typical size are misleading.

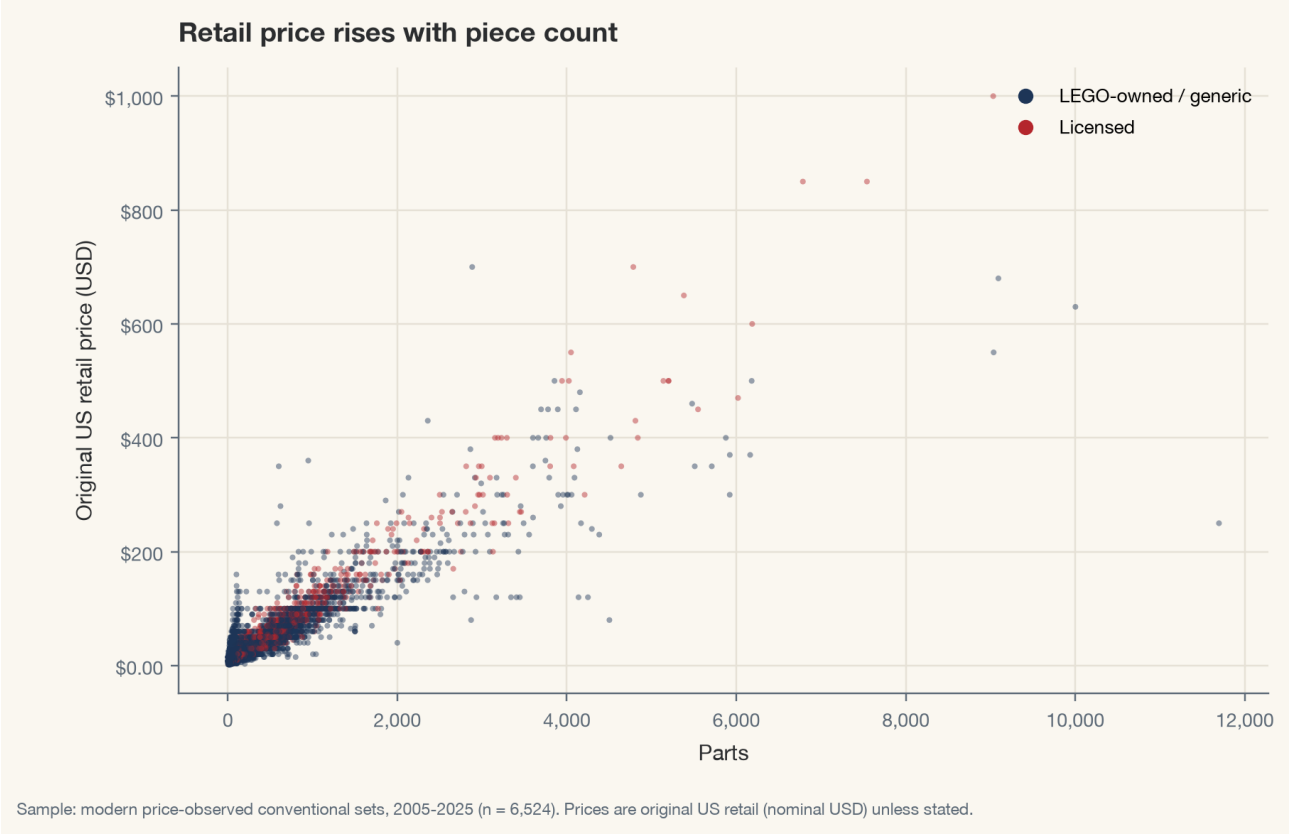


Figure 15. Original US retail price against parts, licensed and LEGO-owned themes.

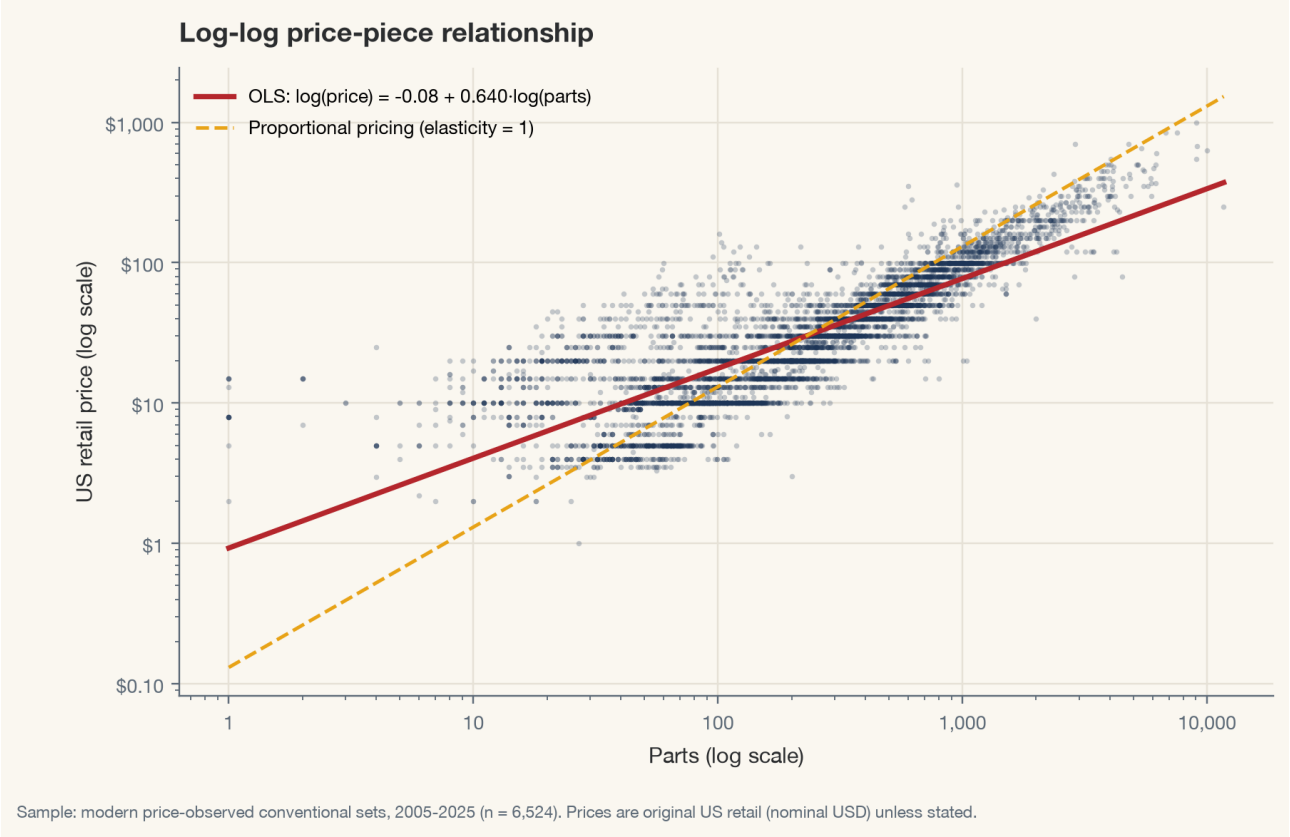


Figure 16. Log-log price-piece relationship with OLS fit and a proportional-pricing reference line.

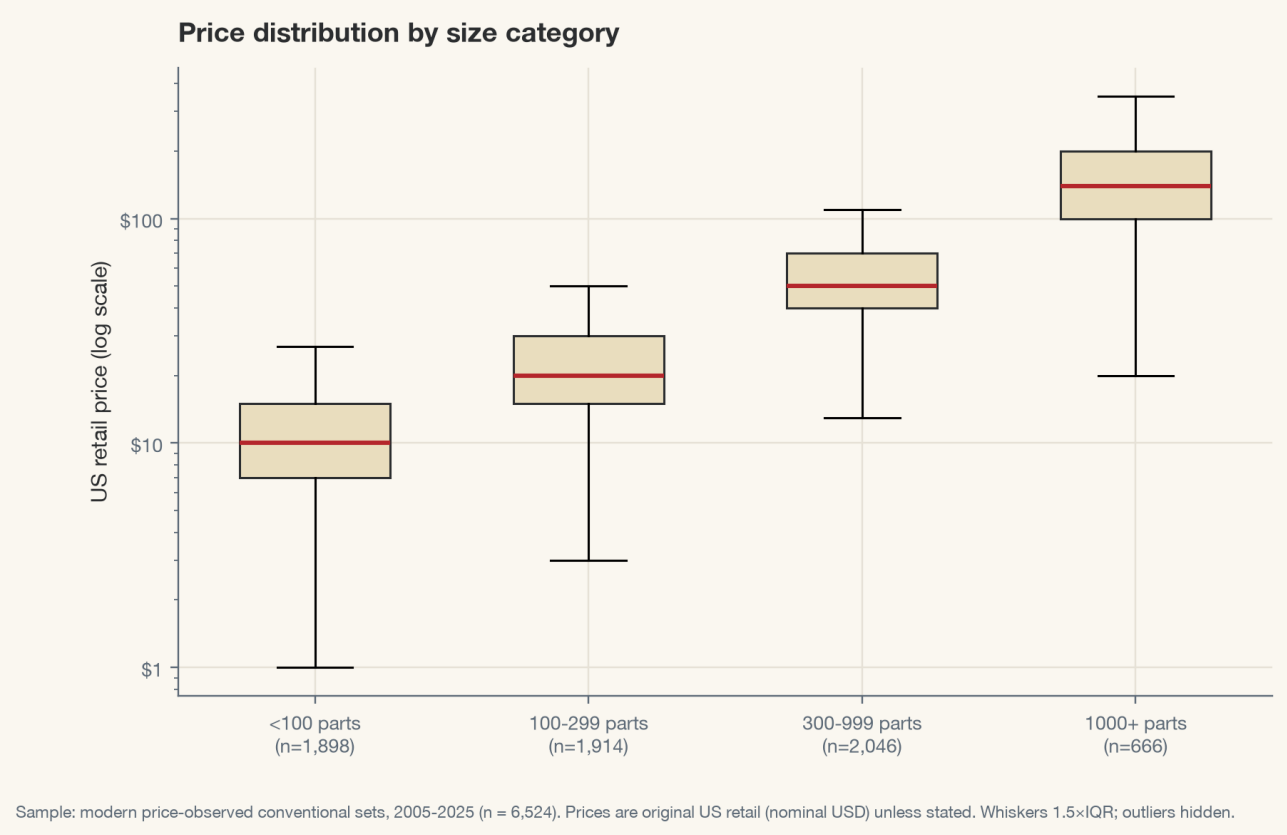


Figure 17. Price distributions by size category.

Inflation adjustment changes the temporal picture (Figure 18). The median nominal price per part was $0.132 in 2005 and $0.107 in 2025; in 2025 dollars it fell from $0.218 to $0.107, a decline of 51%. Set-level Spearman correlations between year and price per part are $\rho = -0.119$ nominal and $\rho = -0.395$ real (both $p < 0.001$). Meanwhile the median nominal set price rose from $15.00 to $49.99 because the typical set became larger. Over the longer priced sample, decade boxplots (Figure 19) show the same downward drift in real price per part, though pre-2005 prices are selectively observed.

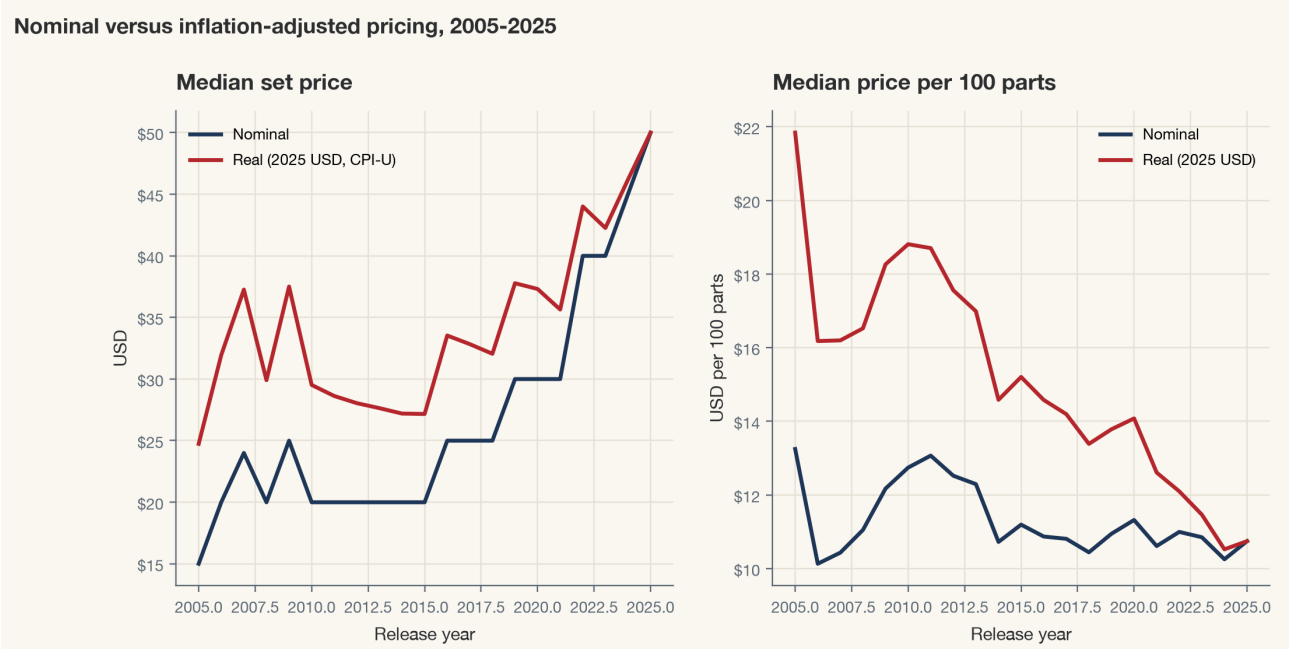


Figure 18. Median set price and price per 100 parts, nominal and in 2025 dollars.

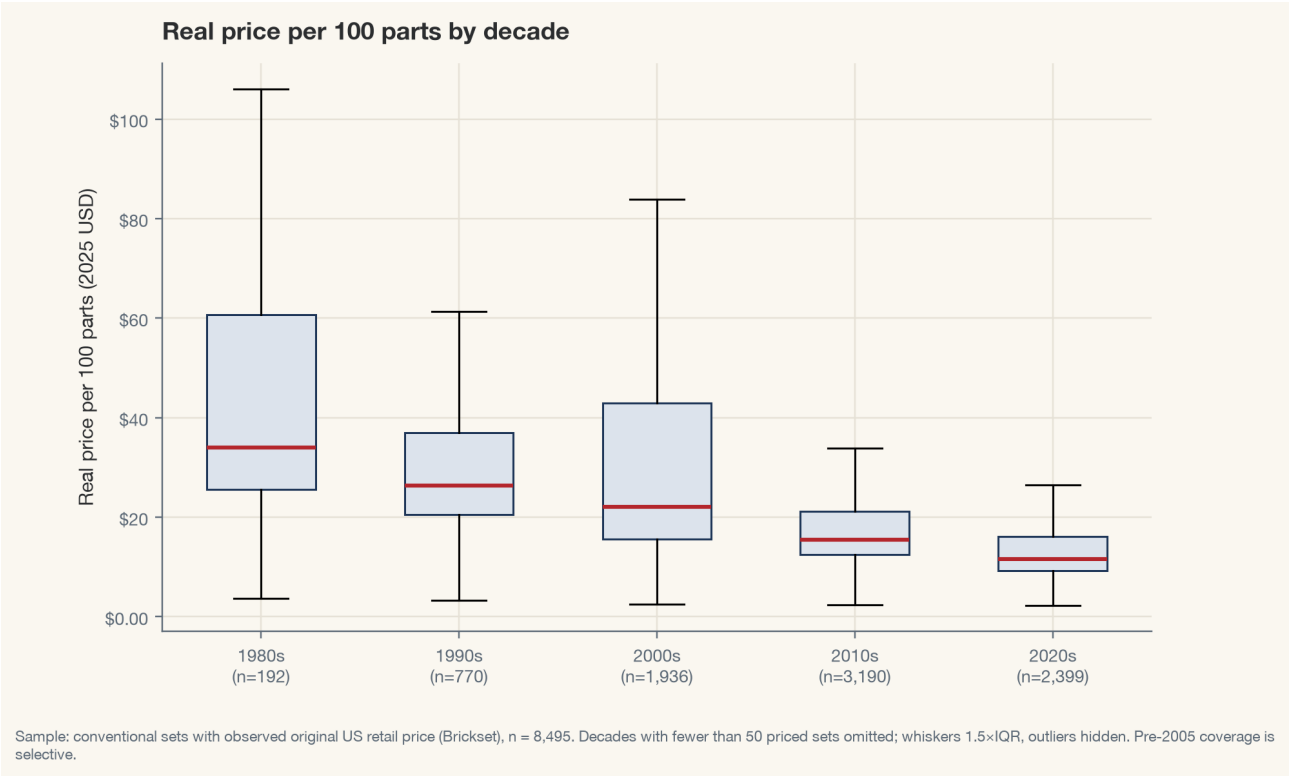


Figure 19. Real price per 100 parts by decade, priced sample.

Unadjusted theme comparisons (Figure 20) place Duplo, Dimensions, Hero Factory and BIONICLE (large-element or figure-heavy lines) at the top of the price-per-part distribution and Friends, Racers, DC and Harry Potter at the bottom. Unadjusted price also rises with minifigure count (Figure 21), but because sets with more figures are larger, these raw patterns do not identify independent associations; Section 6 adjusts for them.

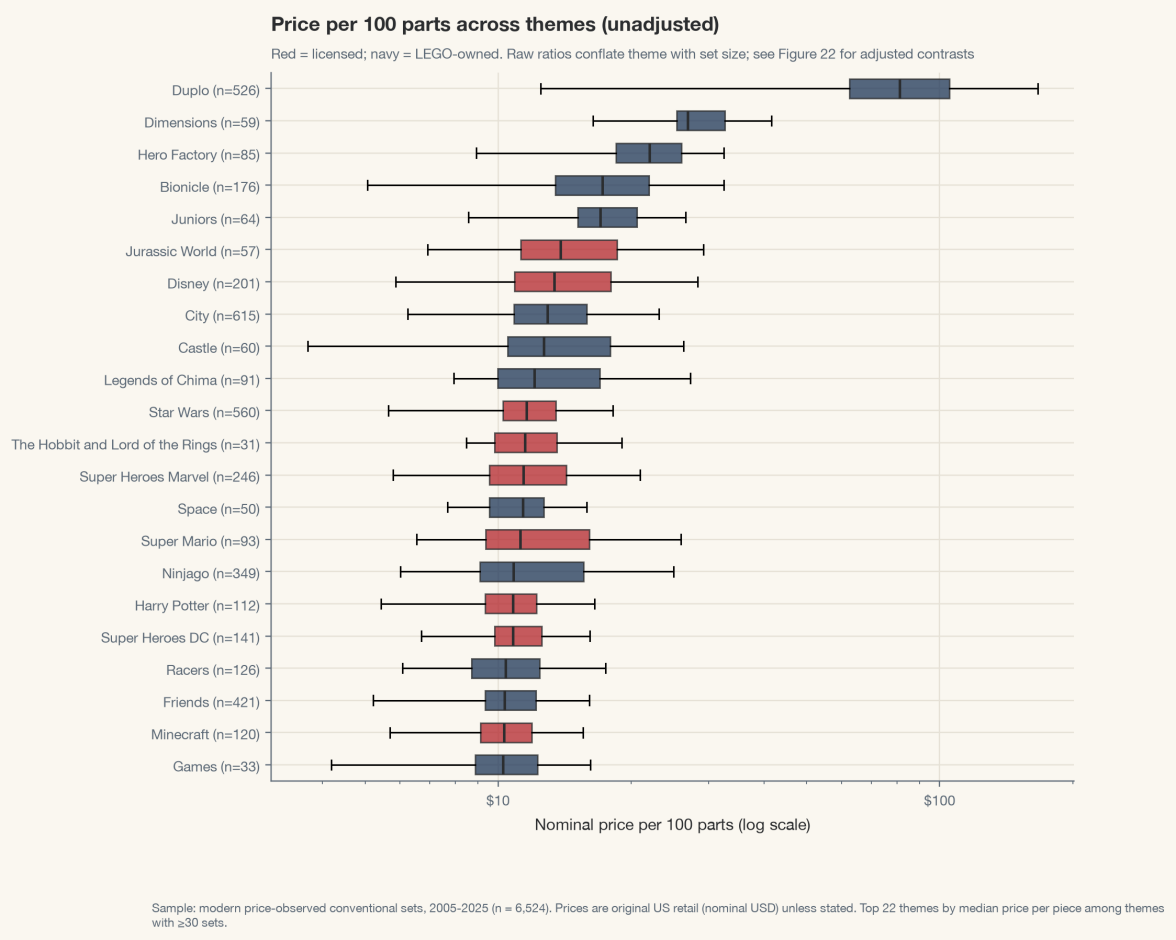


Figure 20. Unadjusted price per 100 parts by theme (log scale).

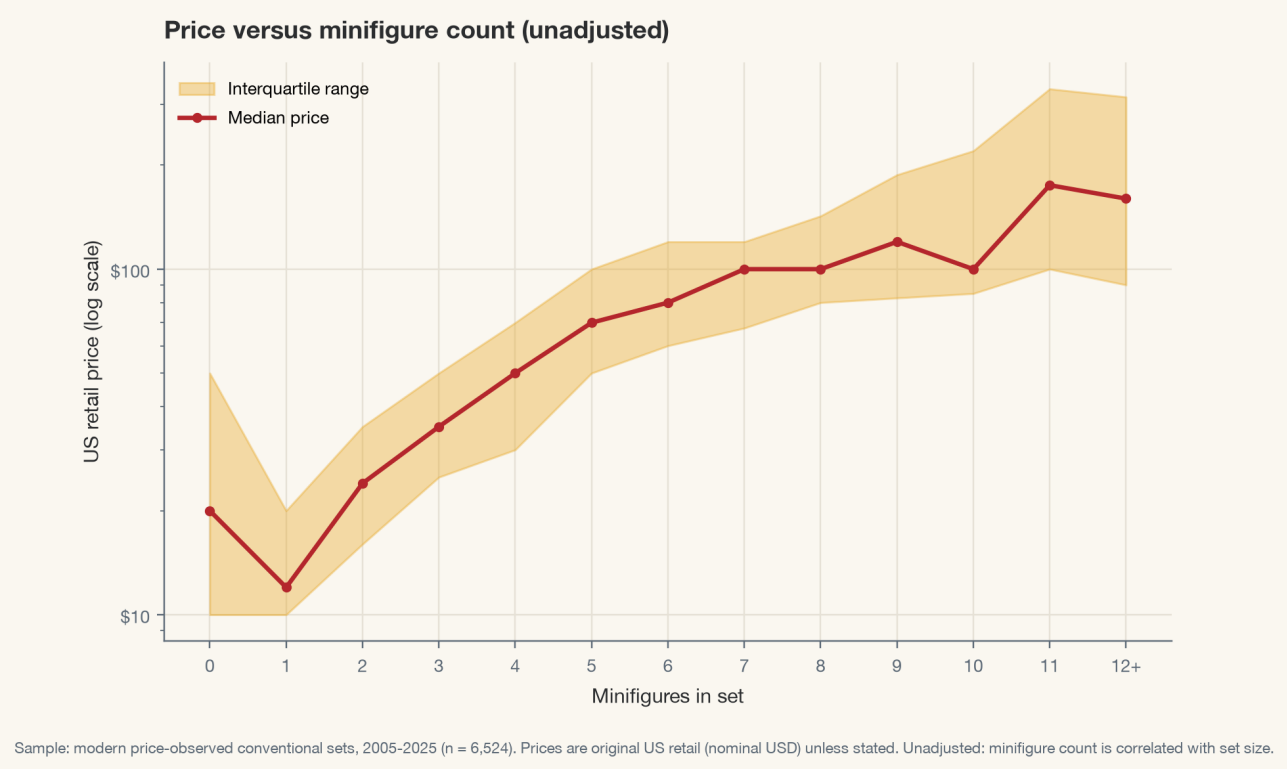


Figure 21. Unadjusted price by minifigure count.

6. Econometric Results

6.1 Size, minifigures, year and licensing

Table 4. Hedonic regression estimates (dependent variable: price in Model 1, log price otherwise). HC3 standard errors in parentheses; one, two and three stars denote p < 0.05, 0.01 and 0.001. Model 4 uses the theme-comparable sample.

Term	(1) Linear	(2) Log-log	(3) Hedonic	(3r) Hedonic, real $	(4) Theme+Year FE	(6) Interactions
Intercept	10.9165***	-0.0754	0.1096*	0.4027***	-0.8216***	0.5795***
	(1.2964)	(0.0467)	(0.0487)	(0.0485)	(0.0898)	(0.1124)
num_parts	0.0821***					
	(0.0033)					
log_parts		0.6400***	0.5852***	0.5827***	0.7194***	0.5054***
		(0.0080)	(0.0090)	(0.0090)	(0.0133)	(0.0192)
mf			0.0449***	0.0457***	0.0467***	0.0384***
			(0.0023)	(0.0023)	(0.0036)	(0.0027)
year_c			0.0085***	-0.0149***		0.0114*
			(0.0013)	(0.0013)		(0.0044)
licensed			-0.0004	-0.0011		-0.0516*
			(0.0115)	(0.0115)		(0.0205)
mf:licensed						0.0195***
						(0.0049)
N	6,524	6,524	6,524	6,524	5,974	6,524
R^2	0.825	0.693	0.707	0.690	0.882	0.711
Adj. R^2	0.825	0.693	0.707	0.690	0.881	0.711
AIC	61,116	10,770	10,482	10,439	4,260	10,399
Theme FE	No	No	No	No	Yes	No
Year FE	No	No	No	No	Yes	No
SE	HC3	HC3	HC3	HC3	HC3	HC3

The linear model (Model 1) attributes $8.21 to each additional 100 parts (95% CI $7.55-$8.86) with $R^2 = 0.825$, but its residuals fan out sharply with size. The log-log model fits better ($R^2 = 0.693$) and yields an elasticity of **0.640** (95% CI 0.624-0.656). Proportional pricing is decisively rejected (Wald F = 2019, p < 0.001): a set with twice as many parts is associated with a price only 56% higher. In the hedonic Model 3 ($R^2 = 0.707$) the elasticity is 0.585, each minifigure is associated with a 4.6% higher price ([4.12%, 5.07%]), and conditional nominal prices drift upward by 0.85% per year ([0.59%, 1.11%]). In real terms the same drift is -1.48% per year ([-1.74%, -1.23%]): holding size, figures and licensing fixed, inflation-adjusted prices fell. The licensed coefficient in Model 3 is essentially zero (-0.0%; theme-clustered CI [-15.9%, 18.8%]). Adding theme and year fixed effects (Model 4) leaves the elasticity at 0.719 and the minifigure association at 4.8% (theme-clustered CI [2.4%, 7.3%]).

Diagnostics support the log specification with robust inference. Variance inflation factors are all below 1.38; heteroskedasticity is present (Breusch-Pagan LM = 1238, p < 0.001), justifying HC3 errors; residuals are heavier-tailed than normal (Jarque-Bera p < 0.001; Figure 22). Dropping the 398 observations with Cook's D > 4/N changes the elasticity to 0.669 and the minifigure coefficient to 0.0485. Figure 23 compares coefficients across specifications.

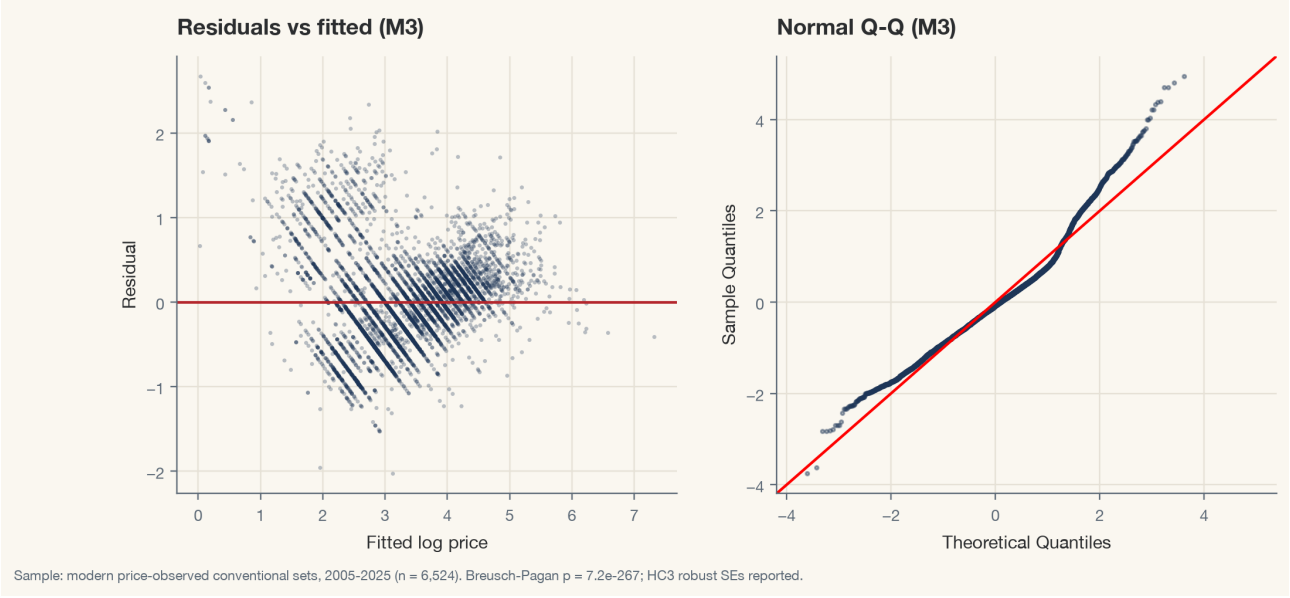


Figure 22. Residual diagnostics for the hedonic model (Model 3).

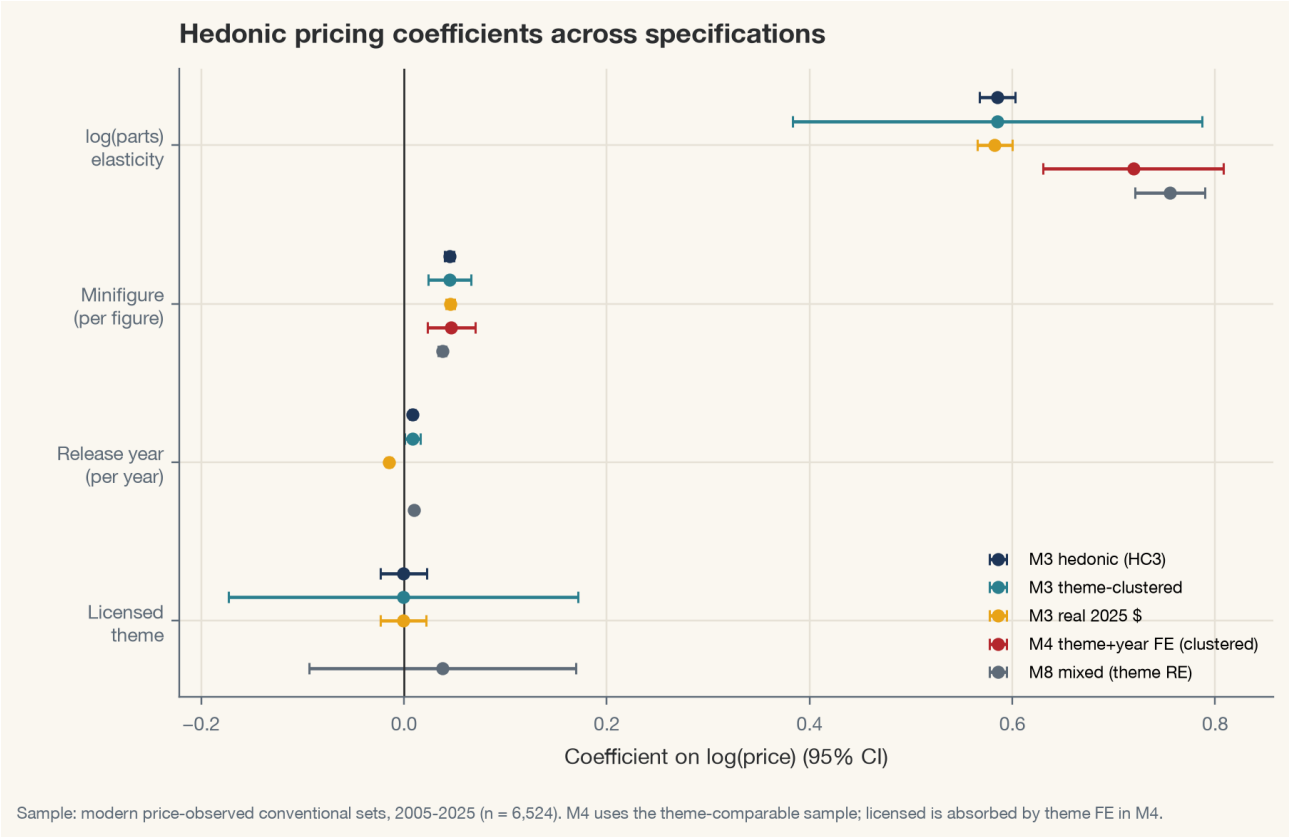


Figure 23. Hedonic coefficients across OLS, clustered, real-price, fixed-effects and mixed-effects specifications.

6.2 Nonlinearity and quantile heterogeneity

Splines improve fit substantially: R^2 rises from 0.707 to 0.787 and AIC falls from 10482 to 8418 (nested F = 222.9 on 11 df, $p < 0.001$). The minifigure spline (Figure 24) is concave: the first few figures carry most of the association, and the curve flattens for large figure counts, where the linear coefficient overstates the conditional difference. Quantile regressions (Figure 25) show that the elasticity is 0.790 at the 10th percentile, 0.728 at the median and 0.487 at the 90th percentile of conditional price; the minifigure coefficient ranges from 0.054 to 0.019, and the licensed coefficient from 0.166 to -0.201. Accumulated local effects from the gradient-boosting model tell a consistent story: the local elasticity is 0.35 over the smallest third of the size range, 0.76 in the middle third and 0.79 in the top third.

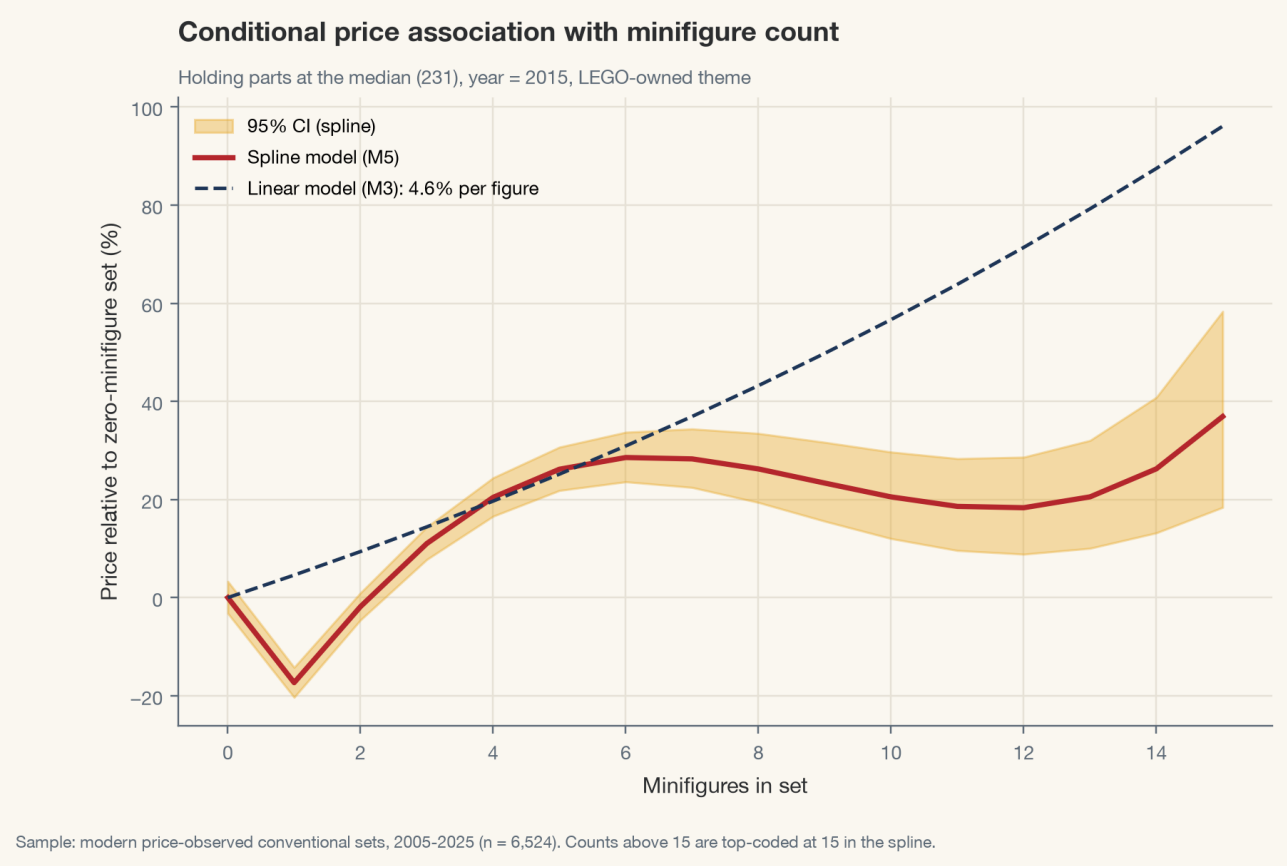


Figure 24. Conditional price association with minifigure count: spline (Model 5) versus linear (Model 3).

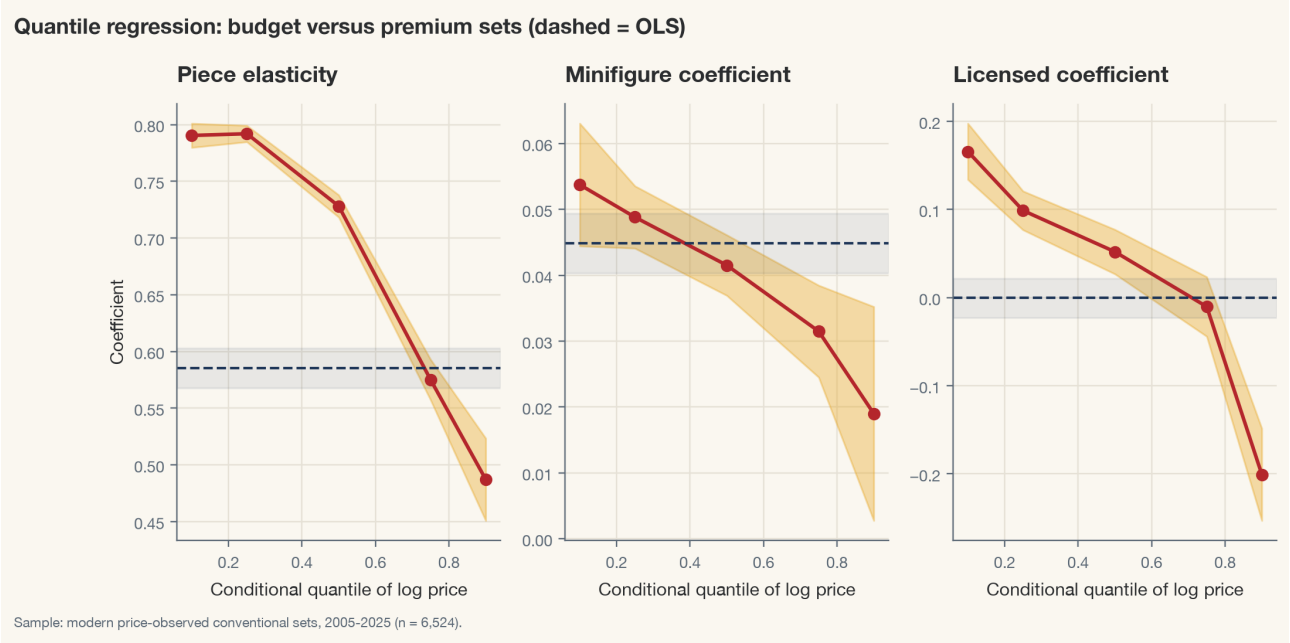


Figure 25. Quantile-regression coefficients for elasticity, minifigures and licensing, with OLS reference.

6.3 Theme contrasts (RQ2, H2)

Theme identity matters a great deal. In the theme-comparable sample (5,974 sets, 41 themes), adding theme fixed effects raises R^2 from 0.699 to 0.882 (nested $F = 229.1$, $p < 0.001$); the joint HC3 Wald test of the 40 contrasts also rejects equality ($p < 0.001$). After Holm correction, 6 themes are priced significantly above City and 22 significantly below (Figure 26). The largest positive contrast is Duplo (+300%, [274%, 328%]), whose large elements make part counts incommensurable with System bricks; the most negative are DOTS (-49%), Mixels and BrickHeadz, small-element or small-format lines. Star Wars (-2.0%, [-6.0%, 2.3%]) is statistically indistinguishable from City; Harry Potter (-14.2%) and Minecraft (-16.3%) are priced below it conditional on parts and figures; Technic is +11.3% ([3.6%, 19.5%]).

Table 5. Highest and lowest adjusted theme contrasts relative to City (Model 4, HC3). The full ranking is in outputs/tables/t32_theme_premium_ranking.csv.

Theme	N	Mean price	Mean parts	Mean figs	Licensed	Contrast %	CI low	CI high	Holm p
Duplo	526	$31.43	44	0.01	no	+300.3	+274.2	+328.2	<0.001
Dimensions	59	$20.36	75	1.14	no	+71.5	+60.4	+83.4	<0.001
Hero Factory	85	$14.57	77	0.40	no	+26.5	+19.1	+34.4	<0.001
Bionicle	176	$16.90	121	0.30	no	+19.1	+11.1	+27.8	<0.001
Juniors	64	$22.01	133	2.16	no	+15.4	+8.0	+23.3	<0.001
Icons	59	$219.40	2474	3.14	no	+13.1	+3.1	+24.1	0.080
Jurassic World	57	$63.53	497	2.77	yes	+12.4	+2.7	+23.1	0.091
Super Mario	93	$51.99	484	0.00	yes	+12.2	+2.3	+23.2	0.105
Creator	305	$44.77	551	0.73	no	-27.0	-31.0	-22.8	<0.001
Exo-Force	31	$25.86	320	2.26	no	-30.0	-36.6	-22.7	<0.001
Speed Champions	80	$29.86	360	1.70	no	-31.7	-35.3	-28.0	<0.001
Seasonal	204	$28.48	301	2.93	no	-36.1	-39.7	-32.3	<0.001
Make & Create	57	$22.66	414	0.70	no	-38.7	-46.5	-29.9	<0.001
Brickheadz	159	$16.37	222	0.00	no	-41.0	-44.1	-37.8	<0.001
Mixels	81	$4.99	62	0.00	no	-48.0	-50.8	-45.1	<0.001
DOTS	69	$14.80	256	0.00	no	-49.1	-54.1	-43.6	<0.001

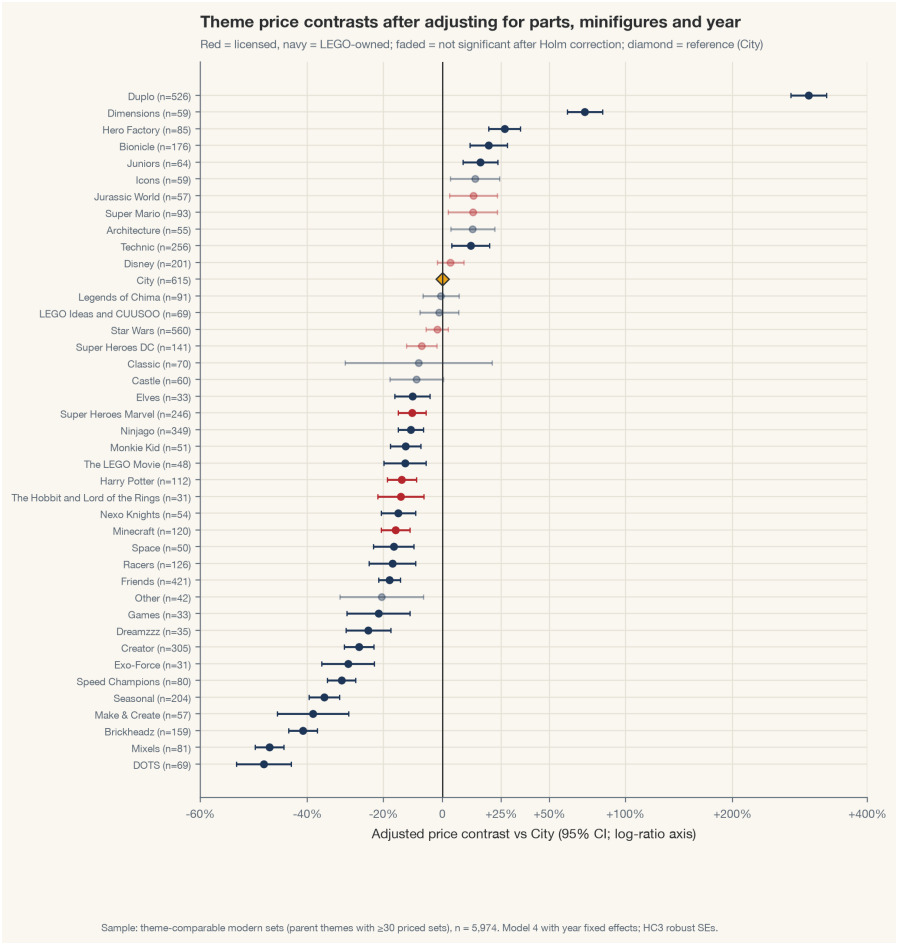


Figure 26. Adjusted theme contrasts relative to City with 95% confidence intervals.

What explains the contrasts? A precision-weighted regression of theme contrasts on theme characteristics (40 themes, $R^2 = 0.32$) finds that themes with larger average sets have lower contrasts ($\beta = -0.383$ per log part, $p < 0.001$), while themes with a higher share of 1,000+ part sets have higher contrasts ($\beta = 1.81$, $p = 0.002$), consistent with a collector/display segment priced above the baseline schedule. Minifigure density ($\beta = 0.124$, $p = 0.163$) and licensing ($\beta = 0.201$, $p = 0.092$) are weaker. Contrast magnitudes are essentially unrelated to theme sample size (correlation 0.11) (Figure 27). Theme rankings are stable over time: across 15 themes estimable in both 2005-2014 and 2015-2025, the Spearman correlation of contrasts is 0.80 ($p < 0.001$), although the theme × post-2015 interaction block is jointly significant ($F = 5.50$, $p < 0.001$), indicating that levels shift even as the ordering persists.

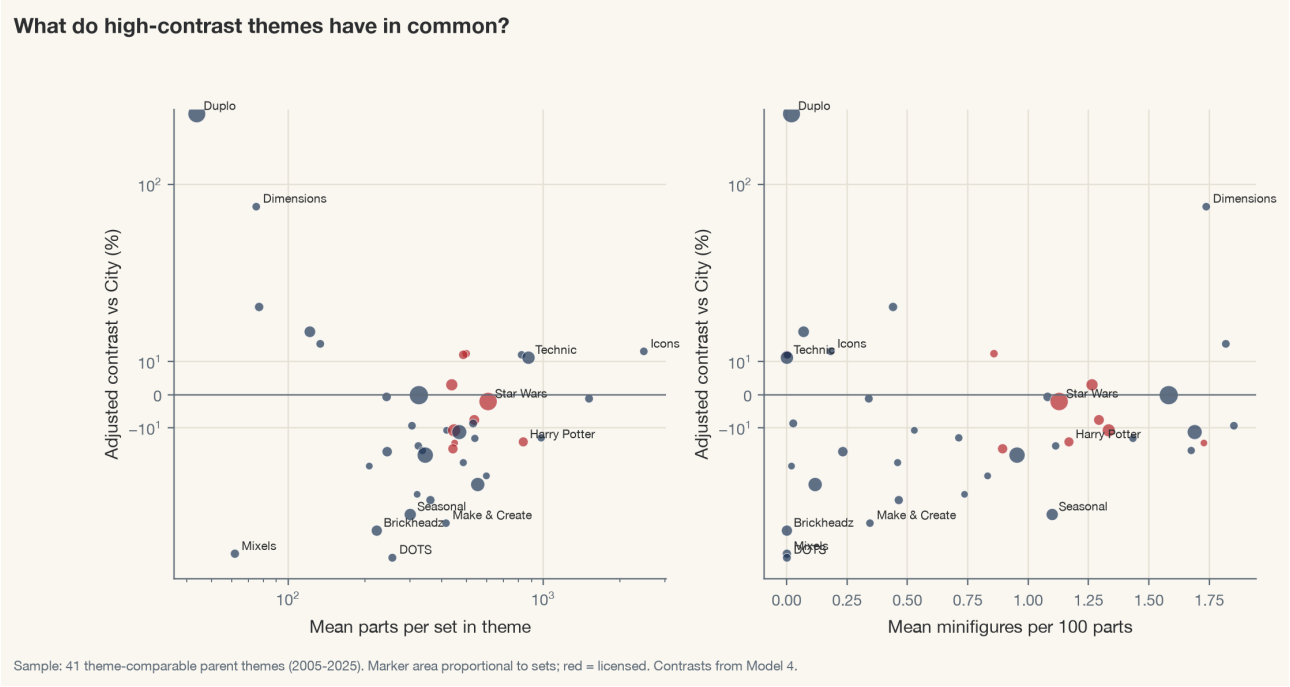


Figure 27. Adjusted theme contrasts against average set size and minifigure density.

The mixed model reaches similar conclusions with partial pooling across all 103 parent themes. The theme random-intercept standard deviation is 0.313 log points, comparable to the residual standard deviation (0.304); in a random-intercept-only model, themes account for 54% of residual variance (ICC). Random slopes for log parts (SD 0.140) improve the log-likelihood by 672 units (likelihood-ratio statistic 1343), so the price-size gradient itself differs by theme. Fixed- and random-effect theme rankings agree closely (Spearman $\rho = 0.91$).

6.4 Licensed themes (H5)

Raw comparisons suggest licensed sets are more expensive: the median licensed set costs $39.99 versus $24.99, but licensed sets are also larger and carry more figures. Conditional estimates are summarised in Figure 28. With controls for parts, minifigures and year, the licensed contrast is -0.0% (theme-clustered CI [-15.9%, 18.8%]); year fixed effects give -0.3% and the theme-cluster bootstrap +2.0% [-14.2%, 20.0%]. These estimates are, however, sensitive to one compositional fact: the LEGO-owned comparison group includes Duplo, whose large elements raise its price per part far above the System schedule. Adding a large-element format control changes the licensed contrast to +10.5% [1.0%, 20.8%], and excluding Duplo, Primo and Quatro gives +10.6% [1.0%, 21.0%]. Propensity matching including the format indicator yields +9.2% [3.6%, 14.7%] for 1,782 matched licensed sets, while inverse-probability weighting gives +1.5% [-14.8%, 20.9%]. The broad classification, which adds ambiguous co-branded themes, yields -6.5% [-24.9%, 16.5%]. By size, the contrast is concentrated in larger sets (+17.5% for 1,000+ parts, [3.8%, 32.9%]), and by era it is negative but imprecise in 2005-2014 and positive but imprecise after 2015. The defensible conclusion is that **licensed System sets are priced roughly 10% above comparable LEGO-owned System sets, but the evidence is fragile**: confidence intervals with theme-level inference are wide, and the estimate depends on how large-element formats are handled. Only 3 of 9 licensed themes in the comparable sample have positive contrasts relative to City.

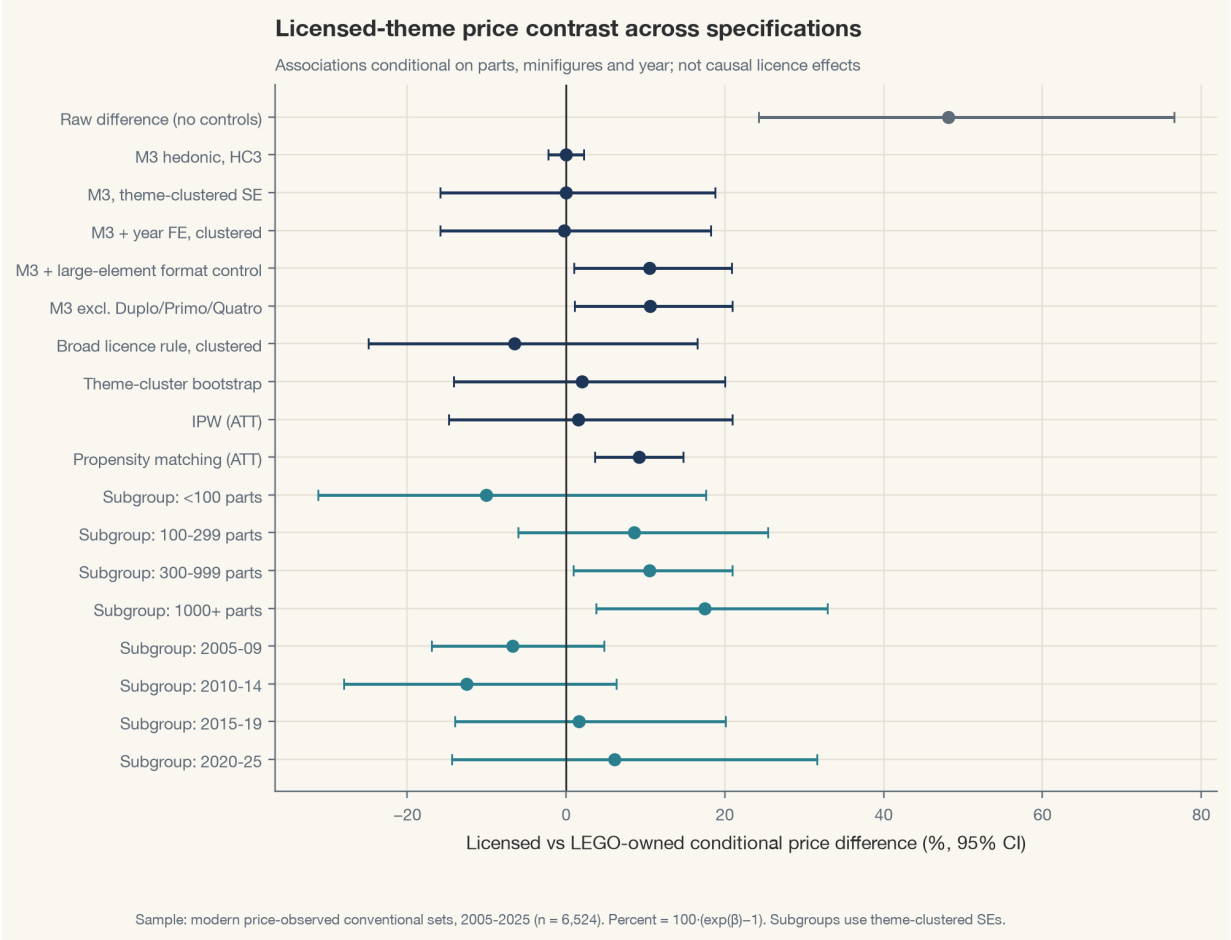


Figure 28. Licensed versus LEGO-owned conditional price contrasts across specifications and subgroups.

6.5 Minifigures (H3)

Minifigures carry a robust independent association with price. Adding minifigure count to the hedonic model is highly significant (nested $F = 249.6$, $p < 0.001$) though it explains only 3.7% of the remaining variance (partial R^2); R^2 rises from 0.696 to 0.707. The pairs-bootstrap interval for the per-figure association is [4.13%, 5.14%]. The association is 5.6% per figure in sets under 300 parts and 2.4% in larger sets, and including any minifigure is associated with a -8.1% higher price. The minifigure × licensed interaction is positive (+1.97% per figure, [1.00%, 2.95%], $p < 0.001$): an additional figure is associated with a larger price difference in licensed themes, where figures depict named characters.

6.6 Temporal change (H4)

The price-size gradient steepened after the first modern era. The elasticity is 0.505 in 2005-2009 and 0.597, 0.601 and 0.602 in subsequent eras (Wald test of equal slopes $F = 20.74$, $p < 0.001$). Rolling three-year estimates (Figure 29) show the rise occurring around 2008-2011 and stability thereafter. The sup-F test locates the single most likely break in the pricing equation in 2022 (sup-F = 47.0, bootstrap 95% critical value 4.9, bootstrap $p = 0.003$; Figure 30), but the elasticity is virtually identical on either side (0.583 before, 0.584 after): the detected break is a shift in the price level rather than in how size is priced. PELT segmentation of annual median set size identifies change points in 1980, 2005, 2015, with segment medians of 74 parts (1975-1979), 48 parts (1980-2004), 108 parts (2005-2014), 213 parts (2015-2025). These dates coincide with the beginning of dense price coverage (2005), so the 2005 break in complexity partly reflects a change in catalogue coverage as well as product change.

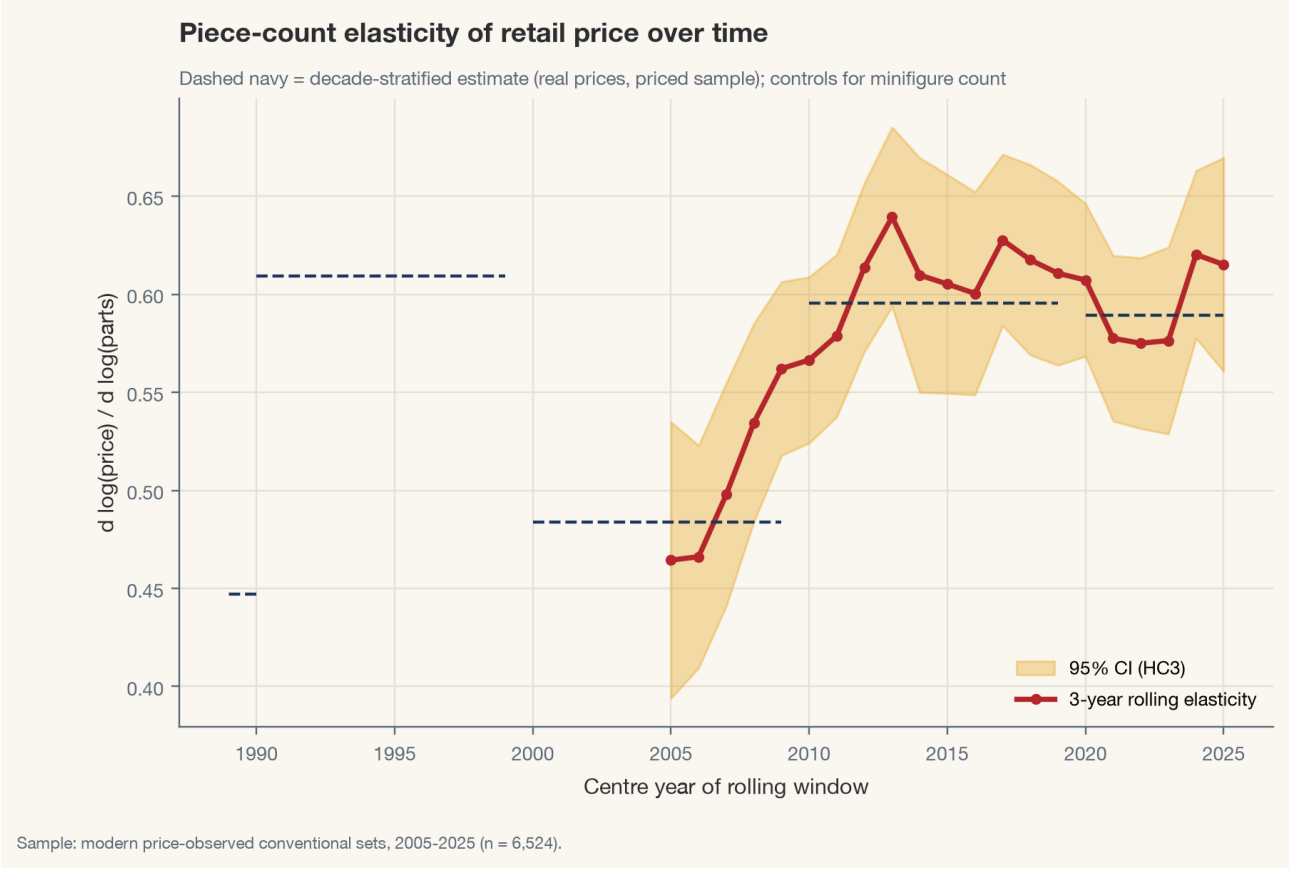


Figure 29. Three-year rolling piece-count elasticity with decade-stratified estimates.

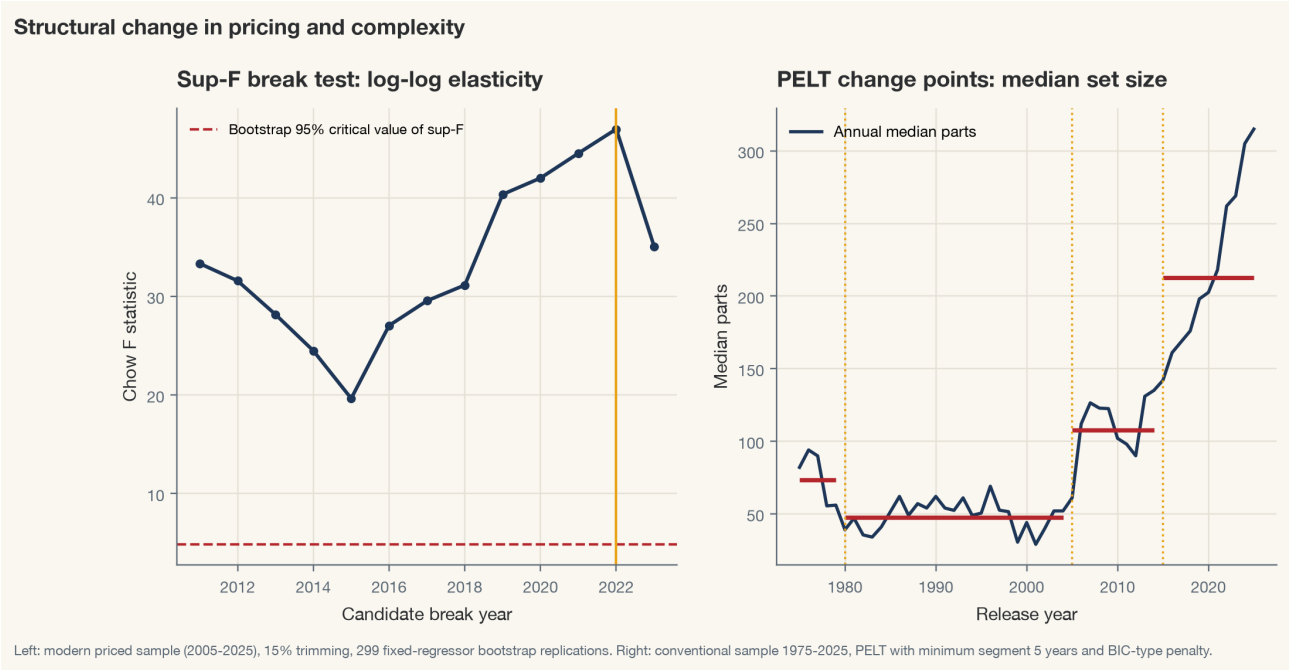


Figure 30. Sup-F statistics by candidate break year (left) and PELT change points in median set size (right).

Decade-stratified real-price regressions on the full priced sample give elasticities of 1980s 0.447 (n = 192), 1990s 0.609 (n = 770), 2000s 0.484 (n = 1,936), 2010s 0.595 (n = 3,190), 2020s 0.589 (n = 2,399). Because pre-2005

prices are observed mainly for larger sets, the earlier estimates should be read as indicative.

6.7 Complexity trend tests (H1)

H1 is rejected by every test. At set level, Spearman ρ between year and parts is 0.384 and Kendall τ 0.264 (both $p < 0.001$). The Mann-Kendall test on annual medians gives $z = 5.57$ ($p < 0.001$) with a Sen slope of 3.32 parts per year; for the annual 90th percentile, $z = 6.97$ and the slope is 15.45 parts per year. Huber robust regression implies growth of 4.71% per year (58% per decade; bootstrap CI for the OLS slope [4.51%, 4.96%]). Median quantile regression gives 5.04% per year and the 90th percentile 3.38%: the middle of the distribution grew proportionally faster than the upper tail, so growth is not driven only by extreme products. With theme fixed effects, within-theme growth is 2.96% per year ([2.14%, 3.78%]), implying that roughly two-thirds of the pooled trend occurs within themes and one-third reflects a shift in theme mix.

Table 6. Summary of formal hypothesis tests.

Hypothesis	Test	Effect size	95% CI	p	N
H1 complexity trend	Spearman rank correlation, year vs parts (set level)	rho = 0.384		<0.001	12,211
H1 complexity trend	Mann-Kendall, annual median parts	Sen slope = 3.32 parts/yr		<0.001	51
H1 complexity trend	Huber robust regression, log parts on year	4.71% per year	OLS bootstrap [4.51%, 4.96%]	<0.001	12,211
H1 complexity trend	Quantile regression q=0.9, log parts on year	3.38% per year	[3.09%, 3.68%]	<0.001	12,211
H2 theme differences	Nested F-test: Model 4 with vs without theme FE	R^2 0.699 -> 0.882		<0.001	5,974
H2 theme differences	Joint HC3 Wald test of 40 theme contrasts	6 positive / 22 negative Holm-significant contrasts		<0.001	5,974
H3 minifigures	Nested F-test: Model 3 with vs without minifigures	4.59% per minifigure	bootstrap [4.13%, 5.14%]	<0.001	6,524
H3 minifigures	Minifigures x licensed interaction (Model 6)	+1.97% per figure in licensed themes	[1.00%, 2.95%]	<0.001	6,524
H4 temporal stability	Wald test, log parts x era interactions (Model 6)	elasticity 2005-09: 0.505, 2010-14: 0.597, 2015-19: 0.601, 2020-25: 0.602		<0.001	6,524
H4 temporal stability	Nested F-test, theme x post-2015 interactions	theme-premium rank correlation early vs late rho = 0.80		<0.001	5,974
H4 temporal stability	Sup-F single-break test (intercept and elasticity), fixed-regressor bootstrap p	break year 2022: level shift; elasticity 0.583 -> 0.584		0.003	6,524
H5 licensed premium	Model 3, theme-clustered SE	-0.0%	[-15.9%, 18.8%]	0.996	6,524
H5 licensed premium	Model 3 + large-element format control, theme-clustered SE	10.5%	[1.0%, 20.8%]	0.030	6,524
H5 licensed premium	Propensity-score matching (ATT)	9.2%	[3.6%, 14.7%]	n/a	1,782

7. Machine Learning Results

7.1 Model comparison

Table 7. Held-out performance in US dollars (log-price models, smeared back-transformation). Random = group-disjoint 20% holdout; temporal = trained 2005-2021, tested 2022-2025.

Model	MAE (random)	RMSE (random)	R^2 (random)	MedAE (random)	MAE (temporal)	RMSE (temporal)	R^2 (temporal)
Extra Trees	$8.06	$18.61	0.916	$3.35	$13.00	$25.61	0.897
XGBoost	$8.22	$20.15	0.902	$3.26	$13.23	$26.74	0.888
HistGradientBoosting	$8.34	$21.21	0.891	$3.37	$14.30	$29.00	0.868
CatBoost	$8.42	$19.29	0.910	$3.52	$13.17	$26.55	0.890
Random Forest	$8.46	$20.90	0.894	$3.25	$12.73	$27.01	0.886
LightGBM	$8.50	$20.45	0.899	$3.46	$13.80	$29.59	0.863
GAM (pygam)	$9.54	$20.90	0.894	$4.32	$15.23	$32.45	0.835
OLS	$10.72	$25.96	0.836	$4.90	$17.51	$48.36	0.633
Ridge	$10.74	$26.03	0.836	$4.83	$17.53	$48.63	0.629
Lasso	$10.93	$26.98	0.823	$4.80	$17.17	$48.29	0.634
Elastic Net	$10.96	$27.05	0.823	$4.81	$17.20	$48.59	0.630
Dummy (median)	$37.67	$64.21	-0.000	$27.28	$44.74	$83.92	-0.104

Every model beats the median baseline (MAE $37.67) by a wide margin (Figure 31). In the random design, **Extra Trees** is most accurate, with MAE $8.06 (bootstrap 95% CI $7.21-$9.03), RMSE $18.61, R^2 0.916 in dollars and 0.948 on the log scale, median absolute error $3.35 and WAPE 16.8%. OLS with theme indicators achieves MAE $10.72 ($R^2$ 0.836), and the GAM $9.54 ($R^2$ 0.894); differences among the tree ensembles are small relative to their bootstrap intervals, so the ranking within that group should not be over-interpreted. The best model reduces MAE by 25% relative to the hedonic linear benchmark. In the temporal design, **Random Forest** is best with MAE $12.73 ([11.63, 13.94]) and R^2 0.886, against $17.51 for OLS. Errors are larger for later years because recent sets are larger and more expensive. Modelling price directly in dollars with LightGBM gives MAE $9.44 in the random design, so the log target with smearing is competitive while guaranteeing positive predictions. Smearing matters: naive exponentiation of log predictions understates prices systematically.

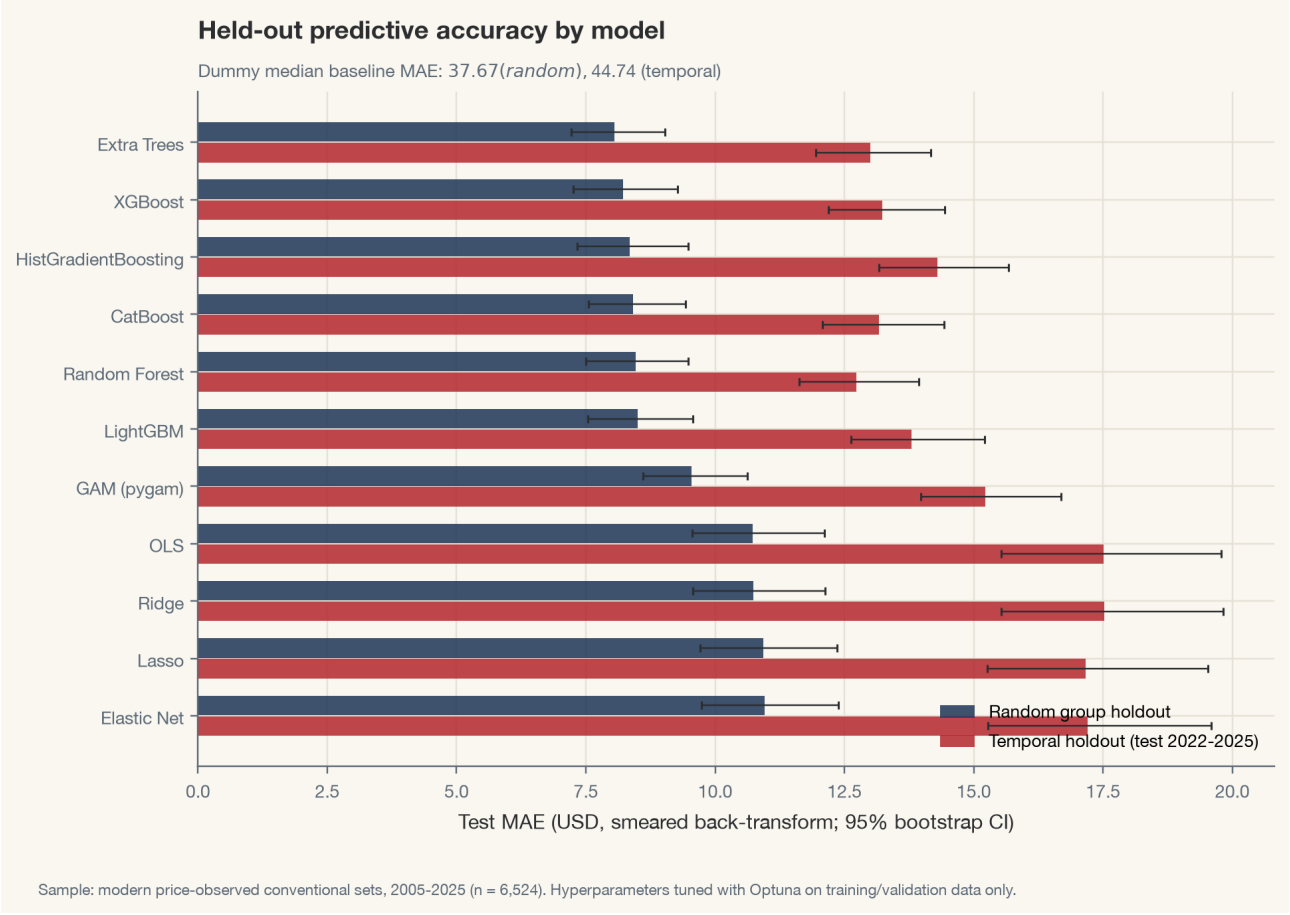


Figure 31. Held-out MAE by model and evaluation design with bootstrap confidence intervals.

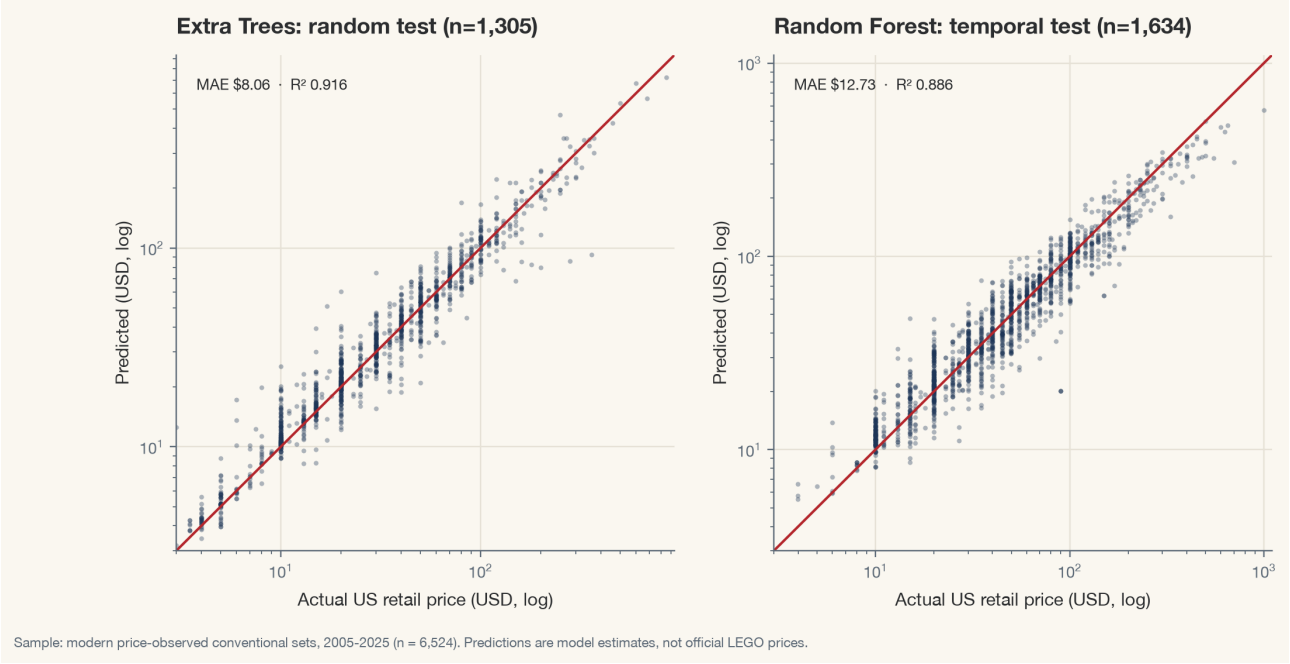


Figure 32. Actual versus predicted prices for the best model in each design.

Predictions track actual prices across three orders of magnitude (Figure 32). Residuals are centred near zero in the random design, whereas the temporal residual median is +0.017 log points (Figure 33), meaning that 2022-2025 sets are priced above what the pre-2022 relationship implies, consistent with the 2022 level shift found by the sup-F test. The rolling-origin evaluation (Figure 34) averages MAE $10.00 for LightGBM versus $13.57 for OLS across test years 2012-2025.

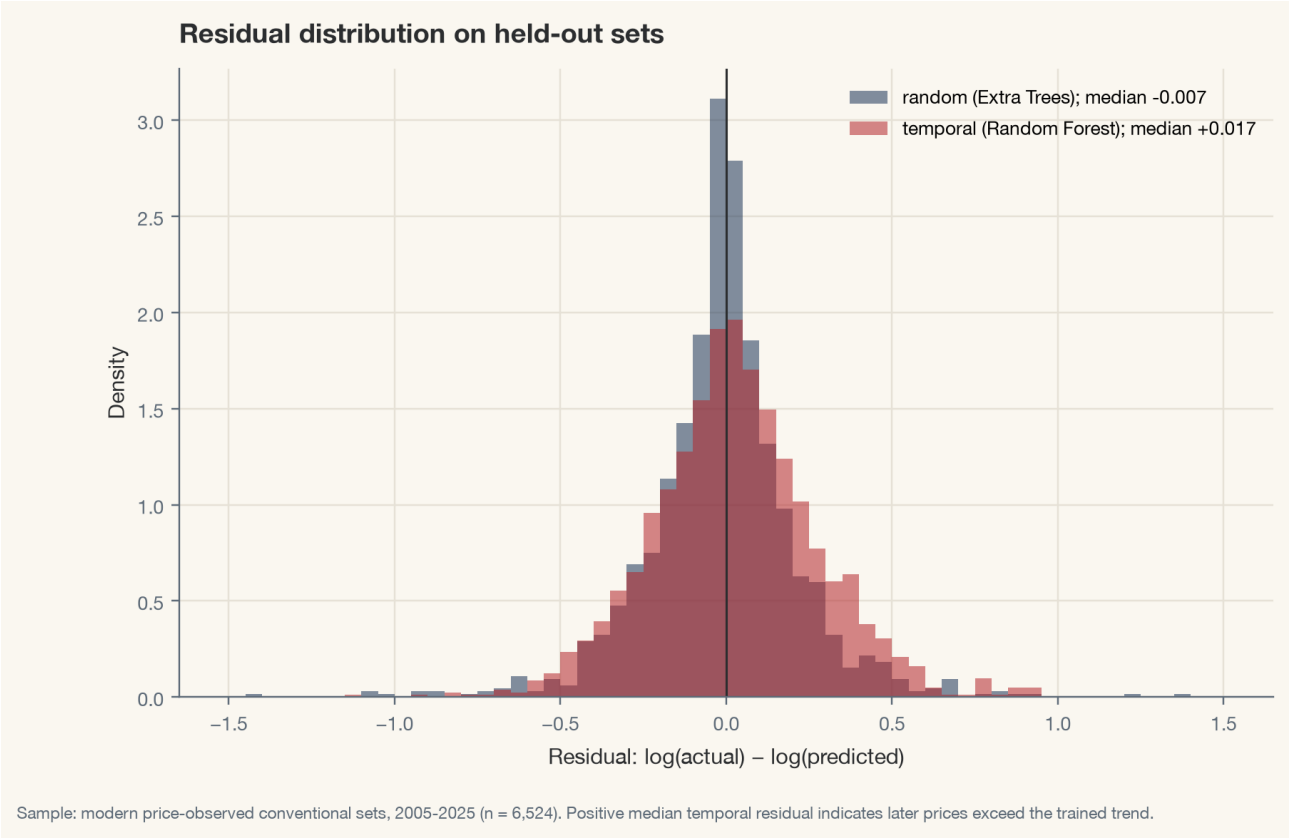


Figure 33. Distribution of log residuals on held-out sets.

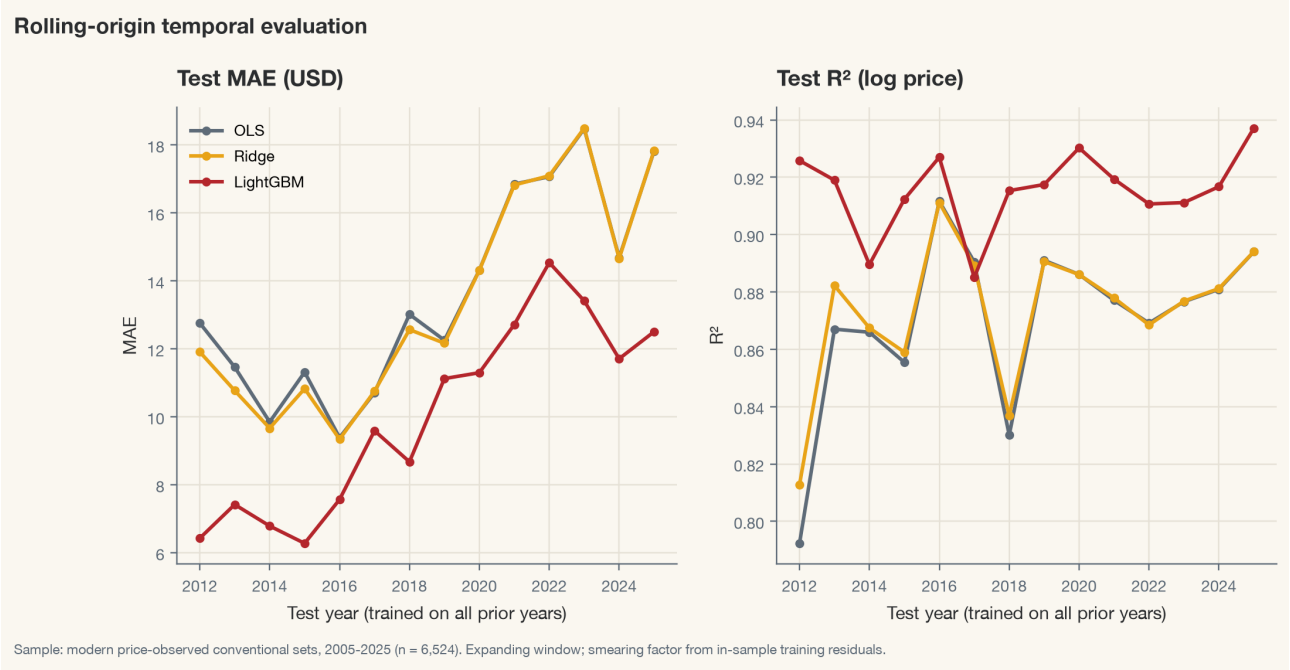


Figure 34. Expanding-window rolling-origin MAE and log-scale R^2 by test year.

Relative errors are largest for the cheapest sets and smallest in the middle of the price range (Figure 35): median absolute percentage error is <$10 8.3%, $10-20 10.8%, $20-40 12.2%, $40-80 12.8%, $80-160 11.8%, $160+ 11.5%. By theme (Figure 36), median errors are largest for Duplo, Technic, Super Mario and Minecraft and smallest for uniform lines such as Mixels, BIONICLE and Ninjago. Medians are slightly over-predicted in every theme because the smearing factor targets the conditional mean, which exceeds the median of a right-skewed distribution. Empirical coverage of an 80% interval built from out-of-fold residual quantiles is 78.8% on the random test set and 89.0% for the 90% interval.

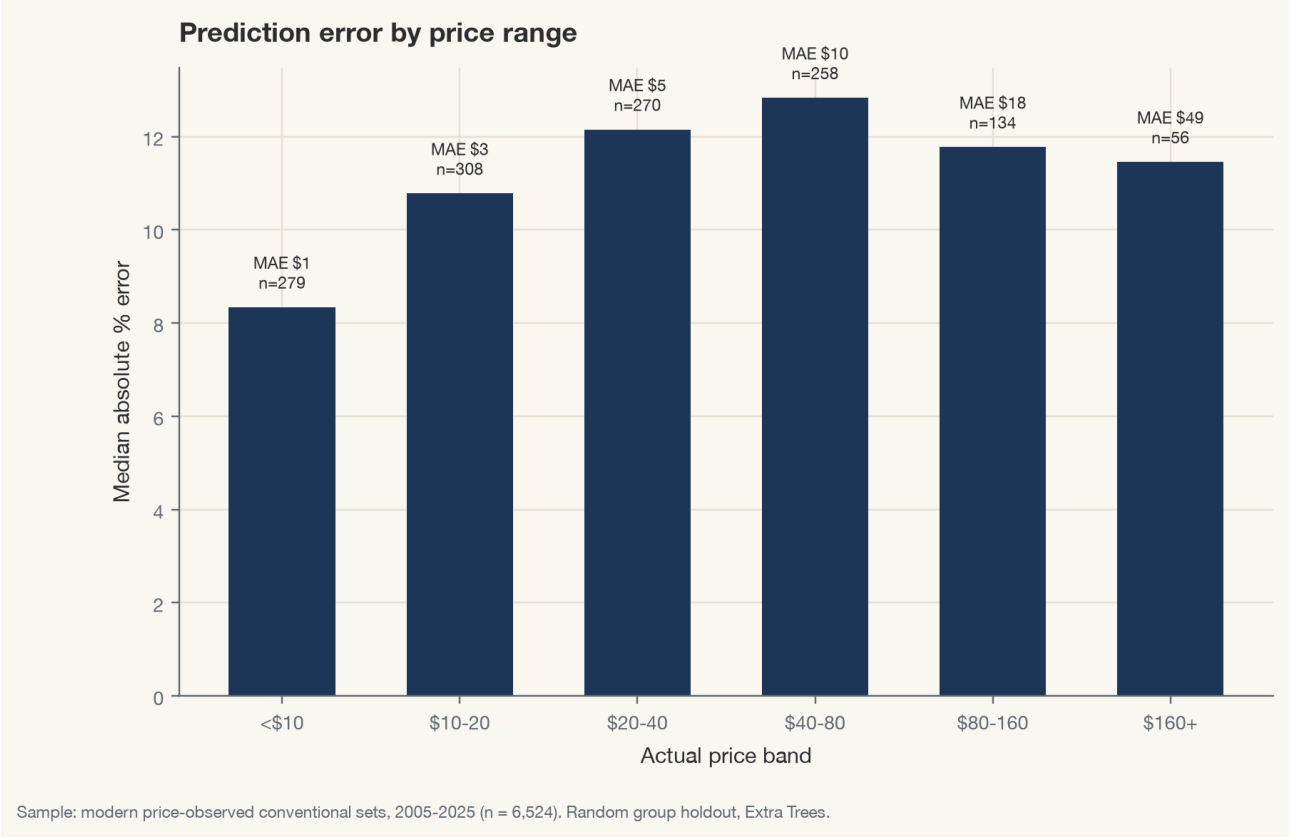


Figure 35. Median absolute percentage error by actual price band.

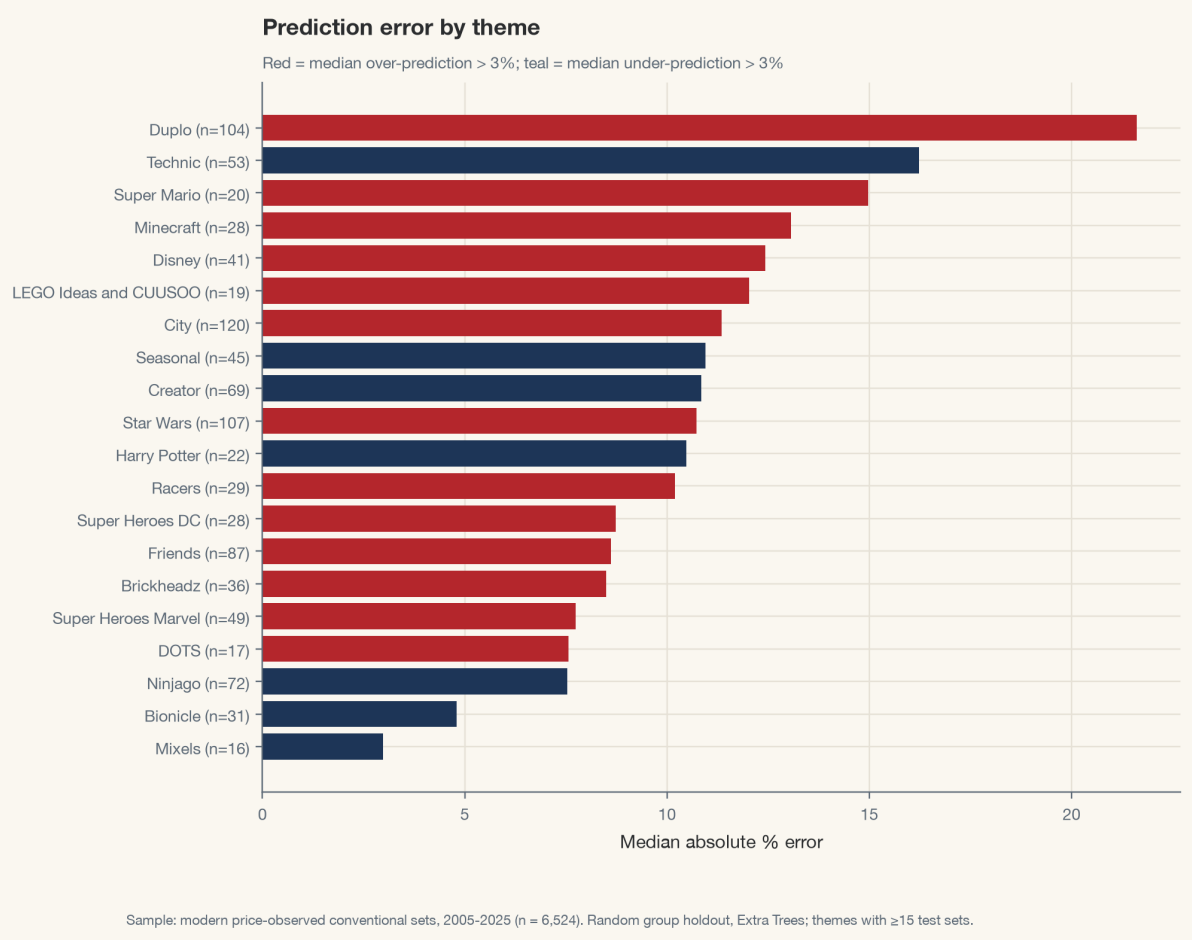


Figure 36. Median absolute percentage error by theme.

7.2 Feature ablation

Table 8. Ablation results: test MAE by information set (random = mean of five group splits; temporal = 2022-2025 test).

Feature set	LightGBM random	LightGBM temporal	OLS random	OLS temporal
A: pieces	$13.74	$18.18	$16.03	$22.86
B: pieces + year	$13.10	$18.39	$16.00	$22.77
C: pieces + year + theme	$8.94	$15.18	$10.89	$16.24
D: pieces + year + minifigs	$11.84	$18.39	$15.64	$20.99
E: all product features	$7.98	$13.81	$10.43	$17.38
F: all excl. theme	$9.00	$14.35	$15.18	$20.77
G: all excl. minifigs	$8.30	$14.51	$10.24	$16.52

The ablations (Figure 37) quantify the predictive contribution of each information set. Piece count alone gives OLS MAE $16.03 and LightGBM $13.74. Adding theme to pieces and year reduces OLS MAE from $16.00 to $10.89 (-31.9%) and LightGBM MAE from $13.10 to $8.94 (-31.8%), improvements that hold in all five splits. Adding minifigures instead changes OLS MAE by -2.2% and LightGBM MAE by -9.7%. Given all other features, removing theme raises LightGBM MAE by 11.4% and removing minifigures by 3.9%; for OLS, removing theme raises MAE by 31.3% while removing minifigures changes MAE by -1.9%. Release year adds little once size is known (-4.6% for LightGBM). Theme is therefore the second most valuable piece of information after size; minifigure counts are statistically significant in hedonic regressions but add only modest predictive value once size and theme are known.

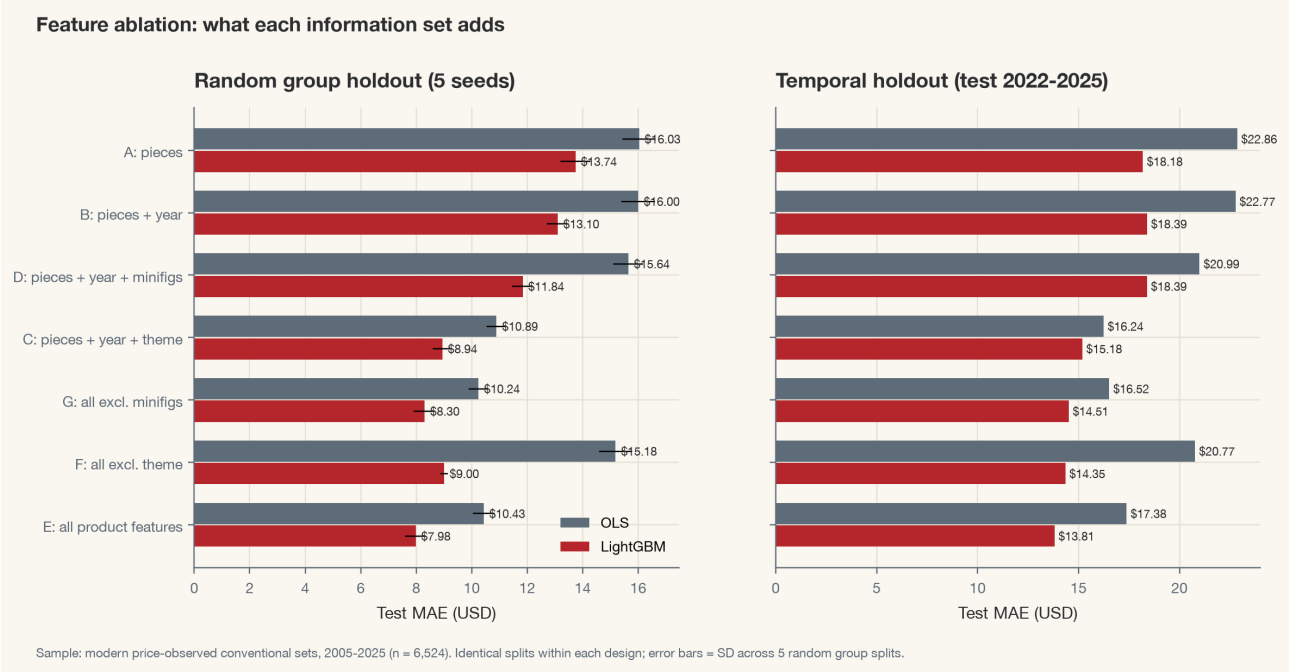


Figure 37. Test MAE for ablation feature sets under random and temporal designs.

7.3 Interpretability

SHAP values (Figure 38) attribute 62.4% to log parts, 14.4% to parent theme, 5.8% to num minifigs, 4.6% to age min, 3.7% to year of mean absolute attribution. Permutation importance on the held-out set agrees on the ordering of the leading features (Figure 39): permuting log parts raises test RMSE by 1.019 log points, theme by 0.244, and minifigures by 0.054. Dependence plots (Figure 40) show a smooth effect of size that steepens beyond small sets, and a year effect that rises from 2005 to about 2011, dips slightly in the mid-2010s and rises again after 2021; ALE estimates imply a total year effect of +26.4% over 2005-2025 holding other features fixed. Partial dependence and ICE curves (Figure 41) confirm that individual sets follow nearly parallel size curves shifted by theme, with no sharp thresholds. The strongest SHAP interactions are log_parts x parent_theme (0.103), log_parts x num_minifigs (0.039), age_min x log_parts (0.037). Because size, figures and theme are correlated, these importances describe how the fitted model uses information and should not be read as causal contributions.

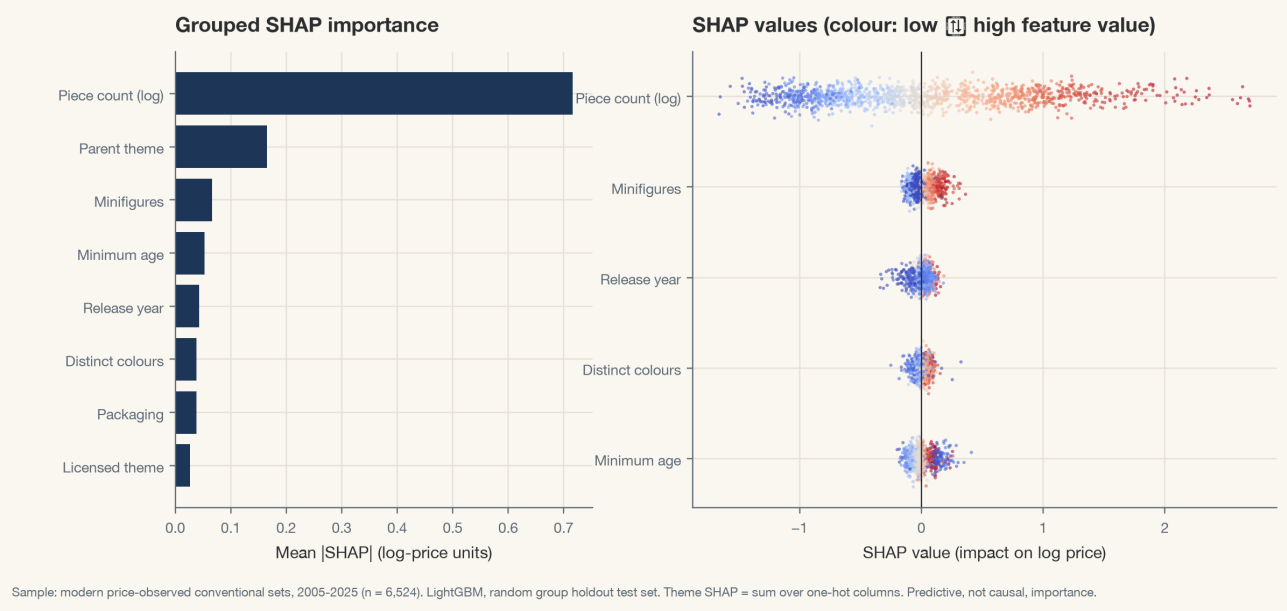


Figure 38. Grouped SHAP importance and SHAP value distributions for numeric features.

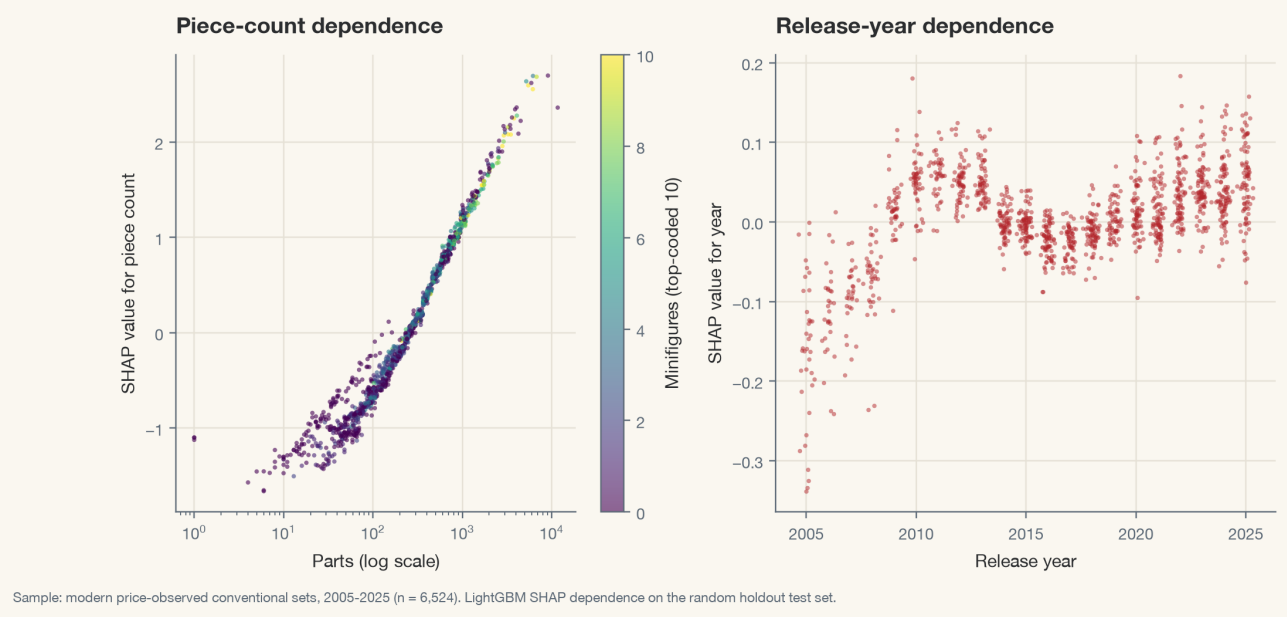


Figure 40. SHAP dependence for piece count and release year.

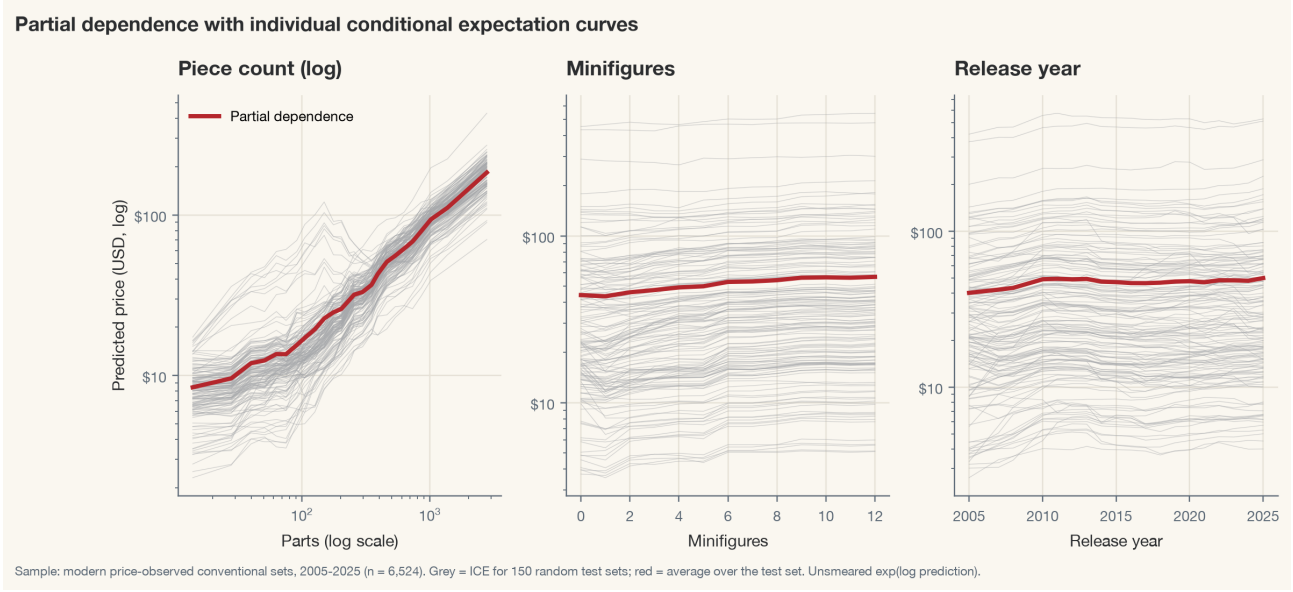


Figure 41. Partial dependence with individual conditional expectation curves.

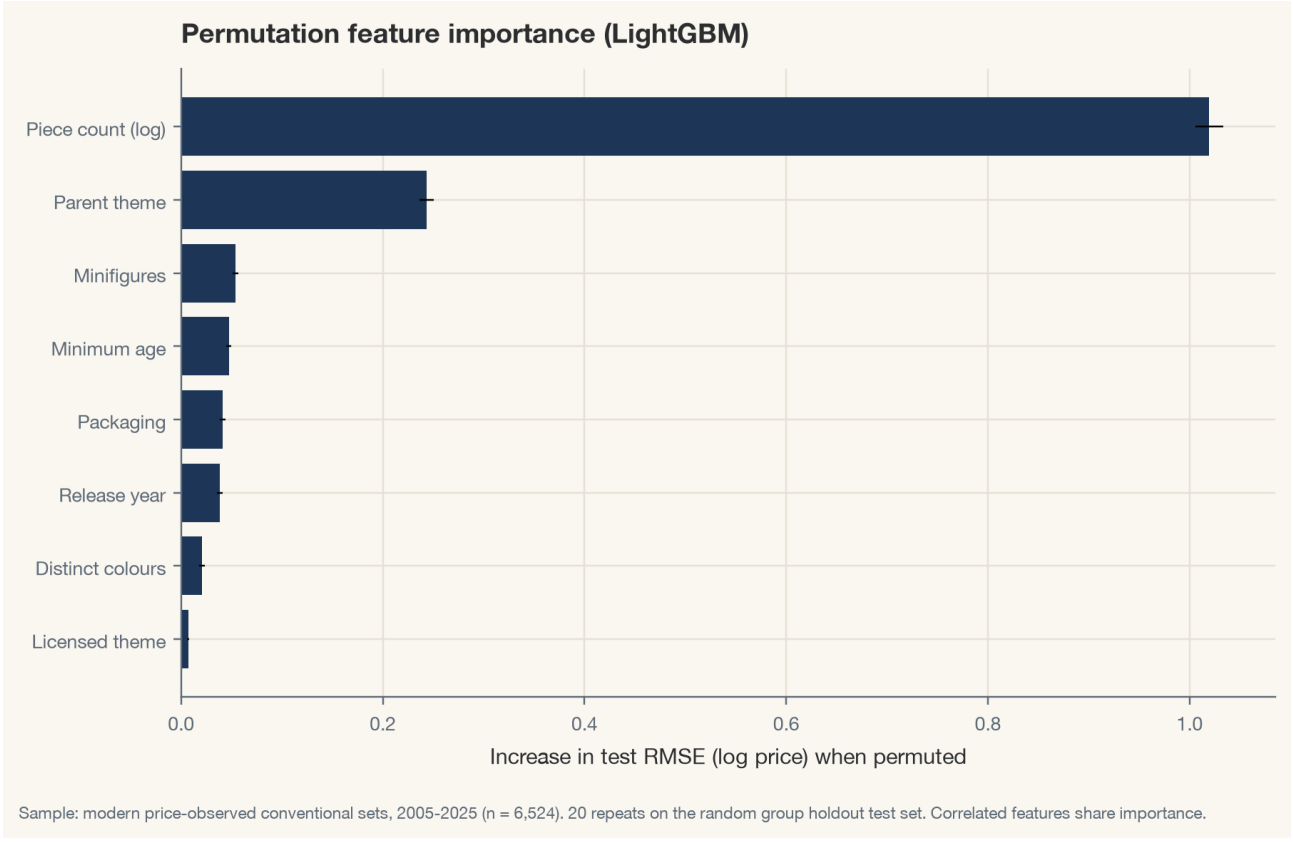


Figure 39. Permutation importance on the held-out test set.

7.4 Minifigure recurrence

Because the release contains no set-minifigure edge list, a bipartite network could not be built without fabricating relationships. The observed recurrence distribution (Figure 42) is informative nonetheless: 80.6% of the 15,790 minifigures appear in exactly one set, only 0.40% appear in ten or more, and the most reused figure appears in 41 sets. Characters are reused mainly by issuing new variants rather than by repeating the same figure; for example, the keyword 'Batman' matches 113 distinct variants. This high exclusivity is consistent with figures functioning as set-specific differentiators.

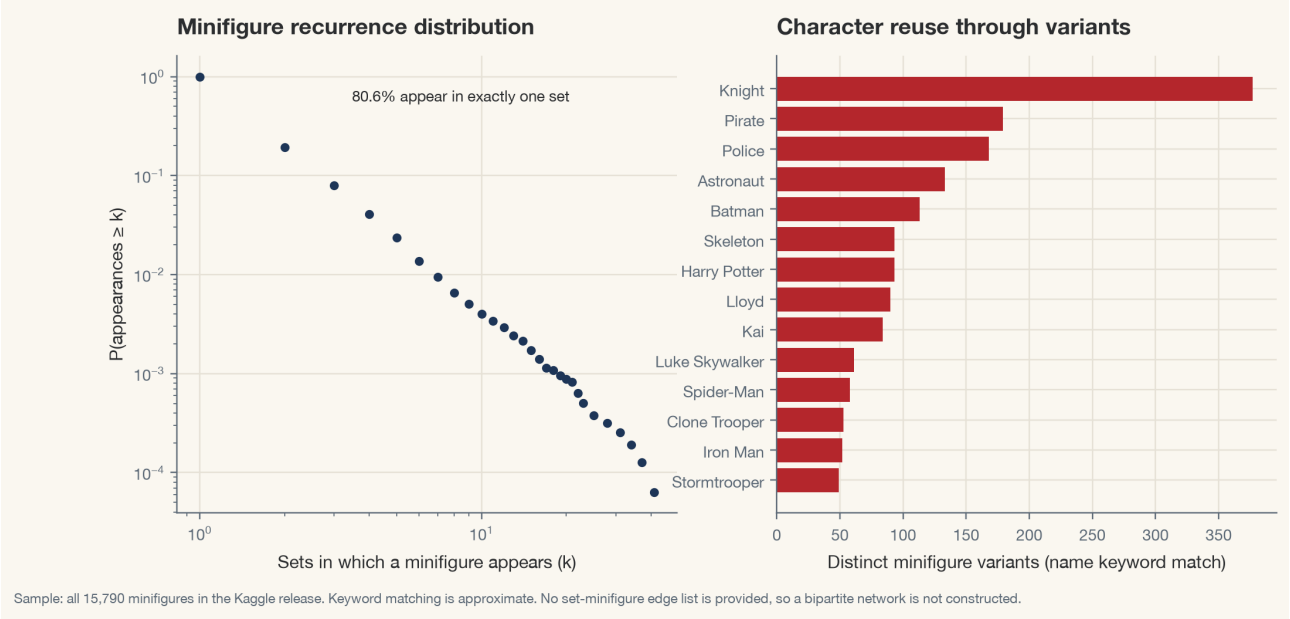


Figure 42. Complementary CDF of minifigure appearances and character variant counts.

8. Robustness Analysis

Table 9 and Figure 43 re-estimate the key findings under 16 pricing specifications (plus four complexity samples), including real prices, alternative standard errors, exclusion of very small and very large sets, recent-decade and theme-comparable samples, Brickset piece counts, archive-only prices, inclusion of nonstandard products and pre-2005 sets, year fixed effects, the broad licensing rule and inverse-probability weights for missing prices.

Table 9. Robustness summary: share of specifications in which each finding keeps its direction, stays within 25% of the baseline magnitude and remains significant at 5%.

Finding	Specs	Direction stable	Within 25% of baseline	Significant	Min	Max
Annual conditional price drift (%)	15	93%	53%	73%	-1.688	1.367
Complexity growth (% per year)	4	100%	75%	100%	2.623	5.247
Licensed contrast (%)	16	100%	81%	75%	5.706	15.193
Minifigure premium (%)	16	100%	94%	100%	4.003	6.851
Piece elasticity	16	100%	94%	100%	0.499	0.788

The piece elasticity is the most stable result in direction and order of magnitude: it is positive, significant and well below one in every specification, ranging from 0.499 to 0.788 (real prices 0.719; Brickset piece counts 0.723; all priced years 0.698). The minifigure association is positive and significant in 100% of specifications (range 4.0% to 6.9%). Complexity growth is positive in all samples (2.62% to 5.25% per year). The nominal conditional price drift is positive, but in real terms it reverses sign (-1.69% per year), which is precisely the nominal-versus-real distinction the analysis is designed to expose. The licensed contrast is the least stable finding: significant in 75% of specifications, within 25% of the baseline in 81%, and ranging from 5.7% to 15.2%; under the broad classification it is +5.7% and including nonstandard products it is +5.7%. Combined with the near-zero estimate obtained without the format control (Section 6.4), this sensitivity is reported rather than resolved.

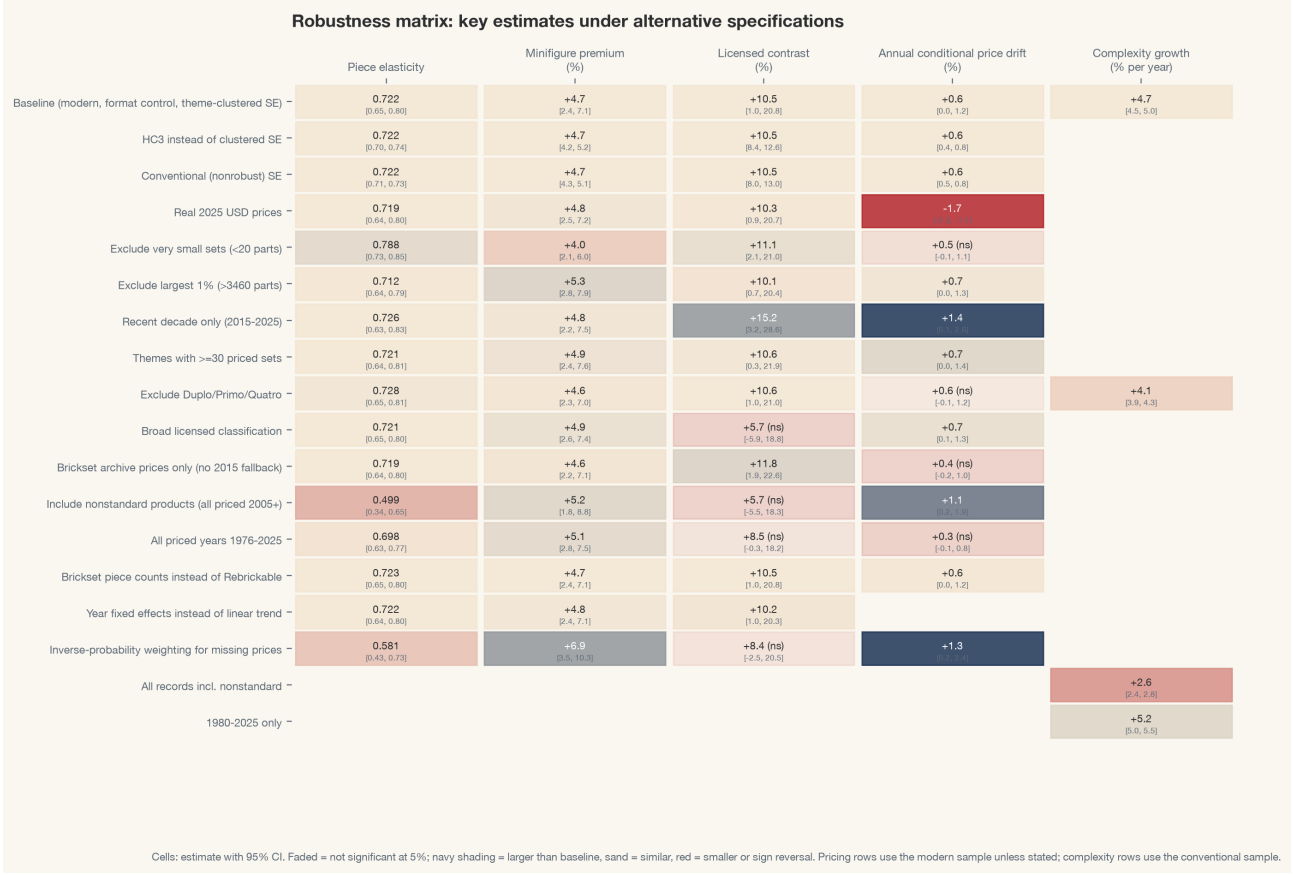


Figure 43. Robustness matrix of key estimates across alternative specifications.

For prediction, alternative temporal boundaries (train through 2015, 2017, 2021 or 2023) produce LightGBM MAE of $10.55 (2016-2019), $11.92 (2018-2021), $13.79 (2022-2025), $12.80 (2024-2025) versus OLS $12.42, $17.66, $17.38, $16.65; LightGBM is more accurate than OLS in every splits.

9. Discussion

What matters most for price. Size dominates. Piece count alone explains 69% of log-price variation in the modern sample, and every interpretability method ranks it first. Theme is clearly second: it lifts the hedonic R^2 by 0.18 and cuts linear-model prediction error by roughly 32%. Minifigures matter conditionally, about 4.6% per figure with a concave profile, but add little predictive accuracy once size and theme are known. Release year matters primarily through inflation and the 2022 level shift.

Sub-proportional pricing. An elasticity near 0.64 means LEGO's price schedule contains a large per-set component that does not scale with parts: boxes, instructions, design, figures and distribution. This makes price per piece a poor value metric across sets of different sizes, since large sets will always look cheap per part. It also reconciles two facts that appear contradictory in popular discussion: median nominal set prices rose markedly while real price per part fell by about half. Both follow from larger sets, an elasticity below one and inflation.

How complexity evolved. Growth in set size is real, sustained and broad-based. It occurs within long-lived themes and is not driven only by the recent wave of very large display sets; the median grew proportionally faster than the 90th percentile. Part of the pooled trend reflects changing theme mix, notably the decline of small basic and town sets and the rise of larger licensed and adult-oriented lines.

How theme differences behave. The largest theme contrasts reflect element format and product segment (Duplo's large bricks; DOTS, Mixels and BrickHeadz small formats; collector-heavy themes) rather than licensing. Once format is controlled, licensed System sets appear modestly more expensive than LEGO-owned ones, but the estimate is fragile. These contrasts are conditional associations: they may reflect licence costs, target-age positioning, element mix (for example printed or specialised parts), or demand differences, none of which the data observe.

What machine learning adds. Gradient-boosted trees reduce error by about 25% relative to the linear hedonic benchmark in the random design and also outperform it out of time. The gain comes from nonlinearity in size and theme-specific size gradients, the same features revealed by the spline and random-slope models. The econometric models remain essential for inference and uncertainty, while ML quantifies the ceiling of predictability from observables.

Product differentiation. The overall picture is of a pricing schedule anchored on physical content with a sizeable fixed component, systematic segment-level offsets that are highly persistent over time, and a catalogue that diversified from a handful of dominant lines to dozens of concurrent themes. The data do not reveal LEGO's internal objectives, and these patterns should not be read as evidence about them.

10. Limitations

- **Observational data.** All estimates are conditional associations. No characteristic is randomly assigned, so coefficients cannot be interpreted as causal effects of adding parts, figures or a licence.
- **Missing and selective prices.** Prices are observed for 69.6% of conventional sets overall and much less before 2005. Inverse-probability weighting does not change the main conclusions, but it corrects only for selection on observables.
- **Historical coverage.** Pre-2005 pricing analyses rely on a 2015 snapshot that over-represents larger and better-documented sets; 1970s estimates are not attempted.
- **Inflation measurement.** CPI-U is a general price index; toy-specific deflators would change the real-price magnitudes, though not the nominal-real distinction.
- **Theme and licence classification.** Licensing is assigned at the parent-theme level with a documented rule. Some LEGO-owned themes contain licensed sets (for example licensed vehicles in Technic and Icons) and vice versa; the broad rule changes the estimate materially.
- **Unobserved costs.** Licence fees, element mix (printed, new or specialised moulds), box size and marketing are unobserved and may drive theme contrasts.
- **No sales volume.** Release counts describe catalogue composition, not sales or popularity; concentration indices measure the release mix only.
- **Data inconsistencies.** Rebrickable and Brickset piece counts agree exactly for only 70% of matched sets, although the median discrepancy is zero and results are robust to the alternative count.
- **Minifigure relationships.** Without a set-minifigure edge list, character-level and network analyses are limited to recurrence counts and keyword matching.

11. Conclusion

Observable product characteristics explain most of the variation in LEGO retail prices. In 2005-2025, piece count alone explains 69% of log-price variance; adding minifigures, year and licensing raises this to 71%, theme fixed effects to 88%, and the best machine-learning model reaches a held-out dollar R^2 of 0.916. The relative contributions have shifted modestly: the price-size elasticity rose from about 0.51 in 2005-2009 to about 0.60 after 2010 and has been stable since, while the most recent years show an upward shift in price level rather than in the size gradient.

The strongest supported findings are:

1. **Complexity growth.** Median conventional set size rose 375% between 1975-1984 and 2016-2025 (CI [283%, 432%]), about 4.7% per year, with roughly two-thirds of the trend occurring within themes.
2. **Sub-proportional pricing.** The log-log elasticity of price with respect to parts is 0.640 (CI 0.624-0.656) and remains well below one in all 16 robustness specifications (range 0.50-0.79; about 0.72 once large-element formats are controlled); doubling parts is associated with a 56% higher price.
3. **Falling real price per part.** Median real price per part fell from $0.218 (2005) to $0.107 (2025) in 2025 dollars; conditional on characteristics, real prices declined 1.48% per year.
4. **Theme is the second pricing dimension.** Adding theme raises R^2 from 0.699 to 0.882 and reduces linear-model MAE by 32%; contrasts are dominated by element format and segment, are stable over time (ρ = 0.80), and a licensed premium of about 10% appears only once format is controlled.
5. **Minifigures matter conditionally, not predictively.** Each figure is associated with 4.6% higher price (CI [4.1%, 5.1%]), with diminishing increments, but removing minifigures from a model that already knows size and theme raises MAE by only 3.9%.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623-2631). https://doi.org/10.1145/3292500.3330701

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica, 61*(4), 821-856. https://doi.org/10.2307/2951764

Apley, D. W., & Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B, 82*(4), 1059-1086. https://doi.org/10.1111/rssb.12377

Bartneck, C., Obaid, M., & Zawieska, K. (2013). Agents with faces: What can we learn from LEGO Minifigures? In *Proceedings of the 1st International Conference on Human-Agent Interaction*. Sapporo, Japan.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Bryer, J. (2026). *brickset: Interface with the Brickset API for getting data about LEGO sets* [R package, data/legosets.rda, commit c411d36]. GitHub. https://github.com/jbryer/brickset

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources, 50*(2), 317-372. https://doi.org/10.3368/jhr.50.2.317

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). https://doi.org/10.1145/2939672.2939785

Court, A. T. (1939). Hedonic price indexes with automotive examples. In *The Dynamics of Automobile Demand* (pp. 99-117). General Motors Corporation.

Dobrynskaya, V., & Kishilova, J. (2022). LEGO: The toy of smart investors. *Research in International Business and Finance, 59*, 101539. https://doi.org/10.1016/j.ribaf.2021.101539

Duan, N. (1983). Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association, 78*(383), 605-610. https://doi.org/10.1080/01621459.1983.10478017

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44-65. https://doi.org/10.1080/10618600.2014.907095

Griliches, Z. (1961). Hedonic price indexes for automobiles: An econometric analysis of quality change. In *The Price Statistics of the Federal Government* (pp. 173-196). National Bureau of Economic Research.

Hacinacar, A. H. (2025). *LEGO sets, themes, and minifigures (1975-2025)* [Data set]. Kaggle. https://www.kaggle.com/datasets/abdurrahmanhakanacar/lego-sets-themes-and-minifigures-1975-2025

Hansen, B. E. (2000). Testing for structural change in conditional models. *Journal of Econometrics, 97*(1), 93-115. https://doi.org/10.1016/S0304-4076(99)00068-8

Hastie, T., & Tibshirani, R. (1990). *Generalized additive models*. Chapman and Hall.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*(282), 457-481. https://doi.org/10.1080/01621459.1958.10501452

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30* (pp. 3146-3154).

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association, 107*(500), 1590-1598. https://doi.org/10.1080/01621459.2012.737745

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33-50. https://doi.org/10.2307/1913643

Kross, S. (2015). *lego: LEGO sets data from Brickset* [R package, data-tidy/legosets.csv, commit 823f2ef]. GitHub. https://github.com/seankross/lego

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*(2), 132-157. https://doi.org/10.1086/259131

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30* (pp. 4765-4774).

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics, 29*(3), 305-325. https://doi.org/10.1016/0304-4076(85)90158-7

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31* (pp. 6638-6648).

Rhoades, S. A. (1993). The Herfindahl-Hirschman index. *Federal Reserve Bulletin, 79*, 188-189.

Robertson, D., & Breen, B. (2013). *Brick by brick: How LEGO rewrote the rules of innovation and conquered the global toy industry*. Crown Business.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34-55. https://doi.org/10.1086/260169

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55. https://doi.org/10.1093/biomet/70.1.41

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379-423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Triplett, J. (2006). *Handbook on hedonic indexes and quality adjustments in price indexes*. OECD Publishing. https://doi.org/10.1787/643587187107

U.S. Bureau of Labor Statistics. (2026). *Consumer Price Index for All Urban Consumers: All items in U.S. city average* [CPIAUCSL], retrieved from FRED, Federal Reserve Bank of St. Louis, 17 November 2026. https://fred.stlouisfed.org/series/CPIAUCSL

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.

Appendix

A. Hyperparameters

Table 10. Optuna-selected hyperparameters.

Design	Model	Trials	Best CV/valid RMSE (log)	Parameters
random	LightGBM	40	0.2278	n_estimators=259; learning_rate=0.1118; num_leaves=22; min_child_samples=5; subsample=0.5127; colsample_bytree=0.9295; reg_lambda=0.2027; subsample_freq=1
random	XGBoost	40	0.2240	n_estimators=475; learning_rate=0.054; max_depth=6; min_child_weight=0.5171; subsample=0.5824; colsample_bytree=0.8734; reg_lambda=1.757
random	HistGradientBoosting	40	0.2266	max_iter=225; learning_rate=0.165; max_leaf_nodes=19; min_samples_leaf=6; l2_regularization=2.882
random	CatBoost	20	0.2279	iterations=1089; learning_rate=0.06492; depth=7; l2_leaf_reg=3.329
random	Random Forest	15	0.2344	n_estimators=280; min_samples_leaf=1; max_features=0.4346
random	Extra Trees	15	0.2286	n_estimators=388; min_samples_leaf=2; max_features=0.762
temporal	LightGBM	40	0.2839	n_estimators=314; learning_rate=0.03899; num_leaves=43; min_child_samples=25; subsample=0.5622; colsample_bytree=0.4401; reg_lambda=0.02679; subsample_freq=1
temporal	XGBoost	40	0.2845	n_estimators=788; learning_rate=0.01058; max_depth=10; min_child_weight=0.6385; subsample=0.7708; colsample_bytree=0.6569; reg_lambda=0.02591
temporal	HistGradientBoosting	40	0.2925	max_iter=772; learning_rate=0.184; max_leaf_nodes=10; min_samples_leaf=50; l2_regularization=0.0007032
temporal	CatBoost	20	0.2872	iterations=575; learning_rate=0.08059; depth=8; l2_leaf_reg=6.137
temporal	Random Forest	15	0.2903	n_estimators=417; min_samples_leaf=1; max_features=0.304
temporal	Extra Trees	15	0.2846	n_estimators=373; min_samples_leaf=1; max_features=0.4234

B. Supplementary regression and quantile results

Table 11. Quantile regression coefficients (Model 7).

Quantile	Term	Coefficient	95% CI
0.1	log_parts	0.7904	[0.7798, 0.8011]
0.1	mf	0.0537	[0.0444, 0.0631]
0.1	year_c	0.0057	[0.0032, 0.0082]
0.1	licensed	0.1658	[0.1339, 0.1977]
0.25	log_parts	0.7920	[0.7848, 0.7992]
0.25	mf	0.0488	[0.0441, 0.0535]
0.25	year_c	0.0025	[0.0008, 0.0042]
0.25	licensed	0.0988	[0.0768, 0.1208]
0.5	log_parts	0.7282	[0.7184, 0.7380]
0.5	mf	0.0415	[0.0369, 0.0461]
0.5	year_c	0.0050	[0.0031, 0.0070]
0.5	licensed	0.0520	[0.0268, 0.0772]
0.75	log_parts	0.5751	[0.5573, 0.5929]
0.75	mf	0.0315	[0.0245, 0.0385]
0.75	year_c	0.0095	[0.0067, 0.0123]
0.75	licensed	-0.0103	[-0.0440, 0.0235]
0.9	log_parts	0.4872	[0.4509, 0.5235]
0.9	mf	0.0190	[0.0028, 0.0352]
0.9	year_c	0.0124	[0.0075, 0.0173]
0.9	licensed	-0.2008	[-0.2530, -0.1486]

C. Licensed-theme classification

Table 12. Parent themes classified as licensed (core rule) or ambiguous (added under the broad rule). All other parent themes are coded LEGO-owned or generic.

Parent theme	Status	Rationale
Angry Birds	licensed	Rovio
Animal Crossing	licensed	Nintendo
Avatar	licensed	20th Century Studios
Avatar: The Last Airbender	licensed	Nickelodeon / Paramount
Ben 10	licensed	Cartoon Network
Bluey	licensed	Ludo Studio / BBC Studios
Brickheadz	ambiguous	Mix of licensed characters and seasonal LEGO-owned figures
Cars	licensed	Disney / Pixar
DC Super Hero Girls	licensed	DC Comics / Warner Bros.
Despicable Me 4	licensed	Illumination / Universal
Dimensions	ambiguous	Video game with mixed licensed and LEGO-owned packs
Discovery	ambiguous	Co-branded with Discovery Channel / NASA
Disney	licensed	The Walt Disney Company
Fortnite	licensed	Epic Games
Gabby's Dollhouse	licensed	DreamWorks Animation
Ghostbusters	licensed	Sony Pictures
Harry Potter	licensed	Warner Bros. Wizarding World
Horizon Adventures	licensed	Sony Interactive Entertainment
Icons	ambiguous	Adult display line containing several licensed vehicles and properties
Indiana Jones	licensed	Lucasfilm
Jurassic World	licensed	Universal Pictures / Amblin
LEGO Ideas and CUUSOO	ambiguous	Fan-designed sets; a minority carry third-party licences
Minecraft	licensed	Mojang / Microsoft
Minions	licensed	Illumination / Universal
Nike	ambiguous	Brand co-promotion
One Piece	licensed	Shueisha / Toei / Netflix
Overwatch	licensed	Blizzard Entertainment
Racers	ambiguous	Some sets carry automotive licences (e.g. Ferrari); most are original
Scooby-Doo	licensed	Warner Bros.
Sonic The Hedgehog	licensed	Sega
Speed Champions	ambiguous	Licensed automotive marques (Ferrari, Porsche, etc.); brand licence rather than entertainment IP
Speed Racer	licensed	Warner Bros.
SpongeBob SquarePants	licensed	Nickelodeon / Paramount
Sports	ambiguous	Mix of licensed leagues (NBA, NHL) and generic sports sets
Star Wars	licensed	Lucasfilm / Disney film franchise
Stranger Things	licensed	Netflix
Studios	ambiguous	Co-branded film-making theme including Steven Spielberg tie-ins
Super Heroes DC	licensed	DC Comics / Warner Bros.
Super Heroes Marvel	licensed	Marvel Entertainment
Super Mario	licensed	Nintendo
Teenage Mutant Ninja Turtles	licensed	Nickelodeon / Paramount
The Hobbit and Lord of the Rings	licensed	Middle-earth Enterprises / Warner Bros.
The LEGO Movie	ambiguous	LEGO-owned characters produced with Warner Bros.; contains licensed DC cameos
The Legend of Zelda	licensed	Nintendo

Parent theme	Status	Rationale
The Lone Ranger	licensed	Disney
The Powerpuff Girls	licensed	Cartoon Network
Trolls: World Tour	licensed	DreamWorks Animation
Wednesday	licensed	MGM / Netflix
Wicked	licensed	Universal Pictures

D. Full robustness matrix

Table 13. All robustness estimates.

Specification	Finding	Estimate	95% CI	p	N
Baseline (modern, format control, theme-clustered SE)	Piece elasticity	0.722	[0.646, 0.798]	<0.001	6,524
Baseline (modern, format control, theme-clustered SE)	Minifigure premium (%)	4.716	[2.404, 7.080]	<0.001	6,524
Baseline (modern, format control, theme-clustered SE)	Licensed contrast (%)	10.470	[0.983, 20.849]	0.030	6,524
Baseline (modern, format control, theme-clustered SE)	Annual conditional price drift (%)	0.638	[0.031, 1.249]	0.039	6,524
HC3 instead of clustered SE	Piece elasticity	0.722	[0.702, 0.742]	<0.001	6,524
HC3 instead of clustered SE	Minifigure premium (%)	4.716	[4.223, 5.212]	<0.001	6,524
HC3 instead of clustered SE	Licensed contrast (%)	10.470	[8.356, 12.626]	<0.001	6,524
HC3 instead of clustered SE	Annual conditional price drift (%)	0.638	[0.437, 0.839]	<0.001	6,524
Conventional (nonrobust) SE	Piece elasticity	0.722	[0.713, 0.731]	<0.001	6,524
Conventional (nonrobust) SE	Minifigure premium (%)	4.716	[4.292, 5.142]	<0.001	6,524
Conventional (nonrobust) SE	Licensed contrast (%)	10.470	[8.020, 12.976]	<0.001	6,524
Conventional (nonrobust) SE	Annual conditional price drift (%)	0.638	[0.461, 0.815]	<0.001	6,524
Real 2025 USD prices	Piece elasticity	0.719	[0.643, 0.795]	<0.001	6,524
Real 2025 USD prices	Minifigure premium (%)	4.796	[2.467, 7.178]	<0.001	6,524
Real 2025 USD prices	Licensed contrast (%)	10.344	[0.911, 20.659]	0.031	6,524
Real 2025 USD prices	Annual conditional price drift (%)	-1.688	[-2.281, -1.091]	<0.001	6,524
Exclude very small sets (<20 parts)	Piece elasticity	0.788	[0.731, 0.845]	<0.001	6,272
Exclude very small sets (<20 parts)	Minifigure premium (%)	4.003	[2.068, 5.974]	<0.001	6,272
Exclude very small sets (<20 parts)	Licensed contrast (%)	11.136	[2.112, 20.958]	0.015	6,272
Exclude very small sets (<20 parts)	Annual conditional price drift (%)	0.467	[-0.141, 1.079]	0.132	6,272
Exclude largest 1% (>3460 parts)	Piece elasticity	0.712	[0.636, 0.789]	<0.001	6,458
Exclude largest 1% (>3460 parts)	Minifigure premium (%)	5.337	[2.825, 7.911]	<0.001	6,458
Exclude largest 1% (>3460 parts)	Licensed contrast (%)	10.075	[0.676, 20.353]	0.035	6,458
Exclude largest 1% (>3460 parts)	Annual conditional price drift (%)	0.651	[0.042, 1.263]	0.036	6,458
Recent decade only (2015-2025)	Piece elasticity	0.726	[0.628, 0.825]	<0.001	4,200
Recent decade only (2015-2025)	Minifigure premium (%)	4.810	[2.198, 7.488]	<0.001	4,200
Recent decade only (2015-2025)	Licensed contrast (%)	15.193	[3.168, 28.619]	0.012	4,200
Recent decade only (2015-2025)	Annual conditional price drift (%)	1.367	[0.136, 2.613]	0.029	4,200
Themes with >=30 priced sets	Piece elasticity	0.721	[0.636, 0.806]	<0.001	5,974
Themes with >=30 priced sets	Minifigure premium (%)	4.923	[2.352, 7.558]	<0.001	5,974
Themes with >=30 priced sets	Licensed contrast (%)	10.566	[0.312, 21.869]	0.043	5,974
Themes with >=30 priced sets	Annual conditional price drift (%)	0.710	[0.047, 1.377]	0.036	5,974
Exclude Duplo/Primo/Quatro	Piece elasticity	0.728	[0.648, 0.808]	<0.001	5,993
Exclude Duplo/Primo/Quatro	Minifigure premium (%)	4.609	[2.257, 7.015]	<0.001	5,993
Exclude Duplo/Primo/Quatro	Licensed contrast (%)	10.550	[1.034, 20.963]	0.029	5,993
Exclude Duplo/Primo/Quatro	Annual conditional price drift (%)	0.567	[-0.100, 1.238]	0.096	5,993
Broad licensed classification	Piece elasticity	0.721	[0.645, 0.797]	<0.001	6,524
Broad licensed classification	Minifigure premium (%)	4.940	[2.571, 7.364]	<0.001	6,524
Broad licensed classification	Licensed contrast (%)	5.706	[-5.935, 18.787]	0.351	6,524
Broad licensed classification	Annual conditional price drift (%)	0.677	[0.095, 1.262]	0.023	6,524
Brickset archive prices only (no 2015 fallback)	Piece elasticity	0.719	[0.639, 0.799]	<0.001	6,169
Brickset archive prices only (no 2015 fallback)	Minifigure premium (%)	4.595	[2.192, 7.055]	<0.001	6,169
Brickset archive prices only (no 2015 fallback)	Licensed contrast (%)	11.796	[1.932, 22.613]	0.018	6,169
Brickset archive prices only (no 2015 fallback)	Annual conditional price drift (%)	0.359	[-0.239, 0.959]	0.240	6,169
Include nonstandard products (all priced 2005+)	Piece elasticity	0.499	[0.345, 0.653]	<0.001	7,210
Include nonstandard products (all priced 2005+)	Minifigure premium (%)	5.241	[1.764, 8.837]	0.003	7,210

Specification	Finding	Estimate	95% CI	p	N
Include nonstandard products (all priced 2005+)	Licensed contrast (%)	5.745	[-5.476, 18.298]	0.329	7,210
Include nonstandard products (all priced 2005+)	Annual conditional price drift (%)	1.057	[0.228, 1.893]	0.012	7,210
All priced years 1976-2025	Piece elasticity	0.698	[0.630, 0.766]	<0.001	8,495
All priced years 1976-2025	Minifigure premium (%)	5.142	[2.821, 7.515]	<0.001	8,495
All priced years 1976-2025	Licensed contrast (%)	8.541	[-0.318, 18.187]	0.059	8,495
All priced years 1976-2025	Annual conditional price drift (%)	0.342	[-0.143, 0.830]	0.167	8,495
Brickset piece counts instead of Rebrickable	Piece elasticity	0.723	[0.646, 0.799]	<0.001	6,523
Brickset piece counts instead of Rebrickable	Minifigure premium (%)	4.717	[2.405, 7.081]	<0.001	6,523
Brickset piece counts instead of Rebrickable	Licensed contrast (%)	10.459	[1.011, 20.791]	0.029	6,523
Brickset piece counts instead of Rebrickable	Annual conditional price drift (%)	0.629	[0.024, 1.238]	0.042	6,523
Year fixed effects instead of linear trend	Piece elasticity	0.722	[0.645, 0.798]	<0.001	6,524
Year fixed effects instead of linear trend	Minifigure premium (%)	4.755	[2.419, 7.143]	<0.001	6,524
Year fixed effects instead of linear trend	Licensed contrast (%)	10.199	[0.954, 20.291]	0.030	6,524
Inverse-probability weighting for missing prices	Piece elasticity	0.581	[0.430, 0.731]	<0.001	6,524
Inverse-probability weighting for missing prices	Minifigure premium (%)	6.851	[3.470, 10.342]	<0.001	6,524
Inverse-probability weighting for missing prices	Licensed contrast (%)	8.405	[-2.502, 20.533]	0.136	6,524
Inverse-probability weighting for missing prices	Annual conditional price drift (%)	1.308	[0.240, 2.387]	0.016	6,524
Baseline (modern, format control, theme-clustered SE)	Complexity growth (% per year)	4.736	[4.508, 4.965]	<0.001	12,211
All records incl. nonstandard	Complexity growth (% per year)	2.623	[2.409, 2.837]	<0.001	18,540
Exclude Duplo/Primo/Quatro	Complexity growth (% per year)	4.108	[3.880, 4.336]	<0.001	10,717
1980-2025 only	Complexity growth (% per year)	5.247	[5.002, 5.493]	<0.001	11,942

E. Cleaning decisions and reproducibility

Raw files are stored unmodified with SHA-256 hashes in `outputs/metrics/schema_report.json`. Every exclusion is listed in Table 2 and `outputs/tables/t02_exclusion_log.csv`; no extreme value was deleted solely for being extreme. The full analysis is reproduced with `python run_pipeline.py` (random seed 20251975). Software: python 3.13.14, pandas 3.1.1, numpy 2.4.6, scipy 1.16.3, statsmodels 0.14.6, scikit-learn 1.9.0, xgboost 3.4.1, lightgbm 4.7.0, catboost 1.2.10, shap 0.52.0, optuna 5.1.0, pygam 0.12.0, matplotlib 3.11.2, seaborn 0.13.2, plotly 7.1.0, streamlit 1.67.0, ruptures 1.1.10, reportlab 4.5.1, pyreadr 0.5.9. The pipeline writes 43 figures (PNG and SVG), 33 tables, trained models and this paper; an automated test suite checks identifier uniqueness, join cardinality, price transformations, leakage, train-test separation, regression consistency, figure existence and agreement between paper numbers and saved results.