# Forecasting Auckland’s Monthly House Price Index (2018–2025): Transparent Time-Series Baselines with Causal Exogenous Screening

**Thandar Kyi [a],**

[a]Master of Analytics, Auckland University of Technology (AUT), Auckland, New Zealand, ygx1540@autuni.ac.nz

**Supervised by Dr. Shu Su**

## Abstract

We forecast Auckland's monthly House Price Index (HPI) using transparent statistical baselines. A causal, month-end imputation policy (no look-ahead) prepares 2018–2025 data. Rolling-origin cross-validation (2018-01 to 2024-08) selects champions per horizon (h=1,3,6,12 months); a strict hold-out (targets 2024-09 to 2025-08) audits accuracy. Non-seasonal ARIMA is best at 1–6 months horizons; at 12-month horizon, the seasonal-naïve benchmark wins. Diebold–Mariano tests show only a marginal short-horizon gain at h=1; Screening finds no robust extra value beyond a pre-registered exogenous set. We recommend ARIMA (h≤6) and seasonal-naïve (h=12) for practitioners. The study offers a reproducible, regime-robust Auckland HPI forecast template.

## 1. Introduction

Auckland's housing market has cycled through rapid booms and slow corrections while remaining persistently unaffordable. Migration pressures, historically low interest rates, and supply constraints underpin these dynamics. Local evidence indicates that buyers form adaptive expectations—extrapolating recent gains—which feeds back into prices and can amplify cycles (Yang, Zhou, & Rehm, 2020). Investor activity strengthens this loop in a supply-constrained setting (Yang & Rehm, 2021). Such feedback makes short-horizon price forecasting both policy-relevant and technically challenging.

Policy and macro settings further shape the market's structure. The Auckland Unitary Plan (AUP) of 2016 upzoned large areas and is causally linked to sustained increases in permitting, altering the supply trajectory and potentially the seasonal and structural properties of price series (Greenaway-McGrevy & Phillips, 2023). Affordability remains strained: Cox (2025) reported that Auckland's median multiple was at 7.7 in 2024, while the Government's Going for Housing Growth programme and related infrastructure-financing updates aim to expand capacity and influence expectations (Te Tūāpapa Kura Kāinga/HUD, 2024, 2025; New Zealand Government—Beehive, 2025; Dentons, 2024). These shifts raise a practical question: which forecasting methods remain reliable as regimes evolve?

Despite a rich local literature on expectations, speculation, and supply, Auckland lacks a horizon-specific, out-of-sample accuracy benchmark for the House Price Index (HPI) that compares transparent statistical baselines —seasonal-naïve (forecast each month as the value from the same month one year earlier) and exponential smoothing/ETS (a state-space error–trend–seasonal model that learns and updates level, trend, and seasonality)— with ARIMA/ARIMAX models that use causally lagged exogenous drivers, and that evaluates differences with Diebold–Mariano tests. International evidence suggests that, on short monthly panels and under regime shifts, well-tuned classical approaches can match or exceed more complex machine-learning methods (Alias, Zainun, & Abdul Rahman, 2016; Zainun, Mohamed Ghazali, & Mohd Sallehudin, 2016; Nunna, Zhou, & Shakya, 2023). Establishing a clear local benchmark therefore fills both a practical need for stakeholders and a methodological gap.

This study builds a reproducible, regime-aware forecasting pipeline for Auckland's monthly HPI to provide a fair, leakage-safe comparison of classical approaches. It first anchors performance with transparent statistical baselines—seasonal-naïve and ETS—then estimates ARIMA and ARIMAX using a compact, causally lagged set of economically motivated regressors (mortgage rate, net migration, dwelling consents, filled jobs). A small set of additional market-activity indicators (OCR, new listings, total stock, sales, days-to-sell, median price) is screened explicitly for leakage and incremental value so that any gains beyond the baselines can be attributed to information that is both timely and causally admissible.

Methodologically, to operationalise this pipeline, we assemble a monthly 2018–2025 panel, apply causal month-end imputation to avoid look-ahead, and select models via rolling-origin cross-validation from 2018-01 to 2024-08 across horizons h={1,3,6,12}. Champions per horizon are selected by log-RMSE and then audited on a strict time-separated hold-out (targets 2024-09 to 2025-08). We report MAE, RMSE, and MAPE on the log scale for selection, provide level-scale metrics for interpretability and assess differences versus the seasonal-naïve using the Diebold–Mariano statistic.

Together, this design yields four contributions. First, it delivers a practitioner-ready accuracy map by horizon for Auckland HPI, showing when simple baselines suffice (e.g., seasonal-naïve at 12 months) and when autoregressive structure adds value (ARIMA at 1–6 months). Second, it provides a documented, reproducible forecasting template—causal month-end imputation, leakage-aware feature timing, rolling-origin cross-validation, and Diebold–Mariano testing—that others can re-run or extend with new data. Third, it offers an explicit screening framework for market-activity indicators (tables and heatmaps) that distinguishes contemporaneous co-movement from causal, lag-based signal, guiding variable inclusion decisions in future ARIMAX variants. Finally, by exporting the exact analysis artefacts (panel, CV metrics, hold-out evaluations, and plots), the study supports auditability and policy use, and creates a baseline against which richer exogenous sets, regime-switching/state-space models, or micro/hedonic integrations can be credibly compared.

For clarity, the House Price Index (HPI) is an aggregate monthly index of price changes; the seasonal-naïve benchmark forecasts each month as the same month one year earlier; ETS denotes the exponential-smoothing family; ARIMA/ARIMAX are autoregressive integrated moving-average models without/with exogenous regressors. The remainder of the paper reviews relevant literature (expectations, speculation, policy, and method evidence), states the research questions and scope, details the research design and methods, and presents findings, discussion, and conclusions with limitations and future work.

## 2. Literature Review

Auckland's housing market exhibits pronounced boom–bust dynamics in which demand pressure interacts with sluggish, capacity-constrained supply. Local evidence indicates that buyers form adaptive expectations—tending to extrapolate recent gains—so expectations and prices reinforce one another and amplify cycles (Yang, Zhou, & Rehm, 2020). Transaction-level work further shows that investor activity can lift prices in a supply-constrained setting, inviting additional speculative entry and strengthening feedback (Yang & Rehm, 2021). These mechanisms imply regime sensitivity: forecasting approaches that perform well in momentum phases can misfire near turning points, so evaluation must span tranquil and turbulent windows and, where possible, incorporate signals tied to credit conditions and market tightness.

Policy has plausibly shifted supply trajectories too. A quasi-experimental study of the 2016 Auckland Unitary Plan (AUP) documents a statistically significant, sustained rise in permitting following large-scale upzoning (Greenaway-McGrevy & Phillips, 2023). Such changes in the supply pipeline can propagate into price dynamics and alter seasonal or structural properties of the time series. For a short-horizon, operational forecasting task, this motivates the use of dwelling consents as a proxy for supply in reduced-form models while acknowledging the risk of structural breaks. Because our sample largely post-dates the AUP (2018–2025), an explicit pre/post policy dummy is less informative than lagging the supply proxy and validating models with rolling-origin procedures that tolerate evolving regimes.

Beyond policy and expectations, Auckland has experienced exuberant episodes identified by diverse detection frameworks—including fundamentals regressions, right-tailed unit-root tests, present-value models, and hedonic methods—before the GFC and again in the mid-2010s (Yang & Zhou, 2024). Present-value approaches are theoretically appealing but data- and assumption-intensive; right-tailed tests are sensitive to sample span and frequency. The practical implication for monthly HPI forecasting is the same: trend breaks and exuberance can bias naïve projections, so robust benchmarks (Seasonal-naïve; ETS) and time-series cross-validation help guard against overfitting particular windows.

Micro-evidence also cautions about composition. Hedonic analyses for Auckland find persistent premia for heritage attributes even after controlling for structure and location (Bade et al., 2020). An aggregate HPI attenuates but does not eliminate composition effects compared to raw medians. This supports the choice of HPI as the target series and justifies using transparent baselines to interpret month-to-month changes.

Against this backdrop, classical time-series methods remain strong candidates for short-horizon forecasting. Across economic series, Seasonal-naïve and ETS offer tough baselines, while ARIMA/SARIMA performs well when autocorrelation and seasonal structure are pronounced. In housing-adjacent applications, ARIMA frequently outperforms double exponential smoothing when serial dependence is strong (Alias, Zainun, & Abdul Rahman, 2016), and it remains a practical baseline for short-term housing demand (Zainun, Mohamed Ghazali, & Mohd Sallehudin, 2016). Comparative studies on modest monthly panels further show that well-tuned statistical baselines can match or exceed complex machine-learning models, particularly when regimes shift and predictors are few or noisy (Nunna, Zhou, & Shakya, 2023). These evaluations converge on good practice: horizon-wise reporting, rolling-origin cross-validation, and explicit comparison to Seasonal-naïve to quantify incremental value.

Where exogenous information is available, parsimonious ARIMAX designs can add predictive power if variables proxy financing conditions and demand/supply momentum. For Auckland, mortgage rates capture credit costs that shape expectations and speculative behaviour (Yang et al., 2020; Yang & Rehm, 2021); net migration tracks population pressure; dwelling consents reflect the post-AUP supply pipeline (Greenaway-McGrevy & Phillips, 2023); and labour-market indicators such as filled jobs approximate income and sentiment. However, monthly samples are short, many market-activity series co-move contemporaneously with prices, and leakage risks are real; causal lags, parsimony, and rigorous validation are therefore essential (Alias et al., 2016; Zainun et al., 2016).

In sum, the literature explains why Auckland prices display momentum and sharp turns (expectations, speculation, slow supply), how policy reforms have reshaped the supply pipeline, and which forecasting practices tend to work on short monthly series. What remains missing is a local, horizon-specific, out-of-sample benchmark for Auckland's monthly HPI that compares Seasonal-naïve, ETS, ARIMA, and a compact ARIMAX with causally lagged drivers; selects champions by cross-validated log-RMSE; and audits gains against Seasonal-naïve using Diebold–Mariano tests. The present study directly addresses this gap by constructing a reproducible pipeline for 2018–2025, aligning variables to month-end with causal imputation, screening additional market-activity indicators for leakage, and reporting a horizon-by-horizon accuracy map that is both practitioner-ready and extensible for future academic work.

## 3. Research Questions and Scoping

The study aims to build a reproducible, leakage-safe forecasting pipeline for Auckland's HPI, benchmarking short-horizon accuracy across Seasonal-naïve, ETS, ARIMA and ARIMAX, selecting horizon-specific champions via rolling-origin cross-validation and auditing them on a strict hold-out, while testing whether a compact set of causally lagged exogenous drivers adds value beyond univariate ARIMA. In doing so, it delivers an empirical benchmark—a horizon-by-horizon accuracy map under a leak-aware protocol—together with a documented methods template (causal imputation, lagged features, rolling-origin CV, target-based test window) that others can reuse or extend, practical guidance on when exogenous signals help versus when simple baselines suffice, and full transparency through export of the analysis panel, accuracy tables, DM tests, and Appendix screening to support replication and review.

### 3.1 Research Question

The study addresses one primary question supported by three sub-questions as follows:

**Main Research Question**

Our main research question, is *How accurately can transparent time-series models—seasonal-naïve, ETS, ARIMA, and ARIMAX—forecast Auckland's monthly House Price Index (HPI) at short horizons (h = 1, 3, 6, 12 months) over 2018–2025?*

**Sub-questions**

The following are our sub research questions:

a) *Do compact, causally lagged exogenous variables that proxy macro/market drivers—mortgage rates, net migration, dwelling consents, filled jobs—improve forecast accuracy over univariate ARIMA once leakage is controlled?*

b) *At which horizons (1, 3, 6, 12) do models most clearly beat the seasonal-naïve baseline?*

c) *Are any additional market-activity indicators (OCR, new listings, total stock, sales, days-to-sell, median price) worth including after screening for leakage and testing out-of-sample?*

### 3.2 Scope

This section clarifies the study's scope by stating what is included (data, models, variables, and evaluation) and what is excluded to avoid bias and ensure interpretability.

**Scope (Inclusions)**

The analysis covers the REINZ Auckland monthly HPI from January 2018 to August 2025. The dependent variable is modelled on the log scale to stabilise variance and to interpret errors proportionally. Forecast horizons of 1, 3, 6 and 12 months ahead are evaluated. Model comparisons include a 12-month Seasonal-naïve baseline, ETS with additive components where sample length permits, ARIMA estimated without exogenous inputs (with a seasonal option considered in cross-validation), and ARIMAX that augments

ARIMA with a compact, causally lagged set of exogenous variables. The pre-registered ARIMAX drivers are the 1-year fixed mortgage rate (RBNZ), net migration (Stats NZ), Auckland residential dwelling consents (Stats NZ), and filled jobs (Stats NZ), each entered with pre-specified causal lags. Additional market-activity variables—OCR, new listings, total stock, sales, days-to-sell, and median price—are assembled and screened for leakage and standalone value; only total stock at lag 1 is retained as a robustness variant and reported in the Appendix rather than in the baseline.

Data preparation aligns all series to month-end and applies causal imputation (same-month-last-year, then past month-of-year median, then forward-fill), generates *_miss flags, and rounds count-type series, with strict prohibition on look-ahead. Model selection uses rolling-origin cross-validation on 2018-01 to 2024-08 (expanding windows) and a strict target-month hold-out from 2024-09 to 2025-08 for final audit. Champions are chosen per horizon by log-scale RMSE in cross-validation; final performance on the hold-out is reported with MAE, RMSE, and MAPE (on log and level scales as appropriate) and evaluated against Seasonal-naïve using Diebold–Mariano tests.

**Scope (Exclusions)**

The project does not estimate micro-level or hedonic models using property attributes or fine geography; the focus is an aggregate index suitable for timely forecasting. It does not pursue structural causal identification; exogenous variables are used strictly for prediction rather than policy inference, so coefficients are not interpreted causally. Complex machine-learning approaches such as deep networks or large ensembles are deliberately set aside to prioritise transparent, auditable baselines that suit a single monthly series subject to regime shifts. No explicit Auckland Unitary Plan dummy is included within 2018–2025, which largely post-dates the reform; instead, potential supply effects are proxied by dwelling consents and assessed via cross-validation and the hold-out test. Real-time data vintages and full density forecasts are outside scope; the primary emphasis is on point-forecast accuracy and model comparison.

Finally, while OCR, new listings, total stock, sales, days-to-sell, and median price are imputed and inspected, most display contemporaneous co-movement with prices and thus pose leakage risk. After lead–lag screening and a small out-of-sample check, none is promoted to the baseline ARIMAX; only total stock at lag 1 is retained as an Appendix robustness item (Tables A1–A2).

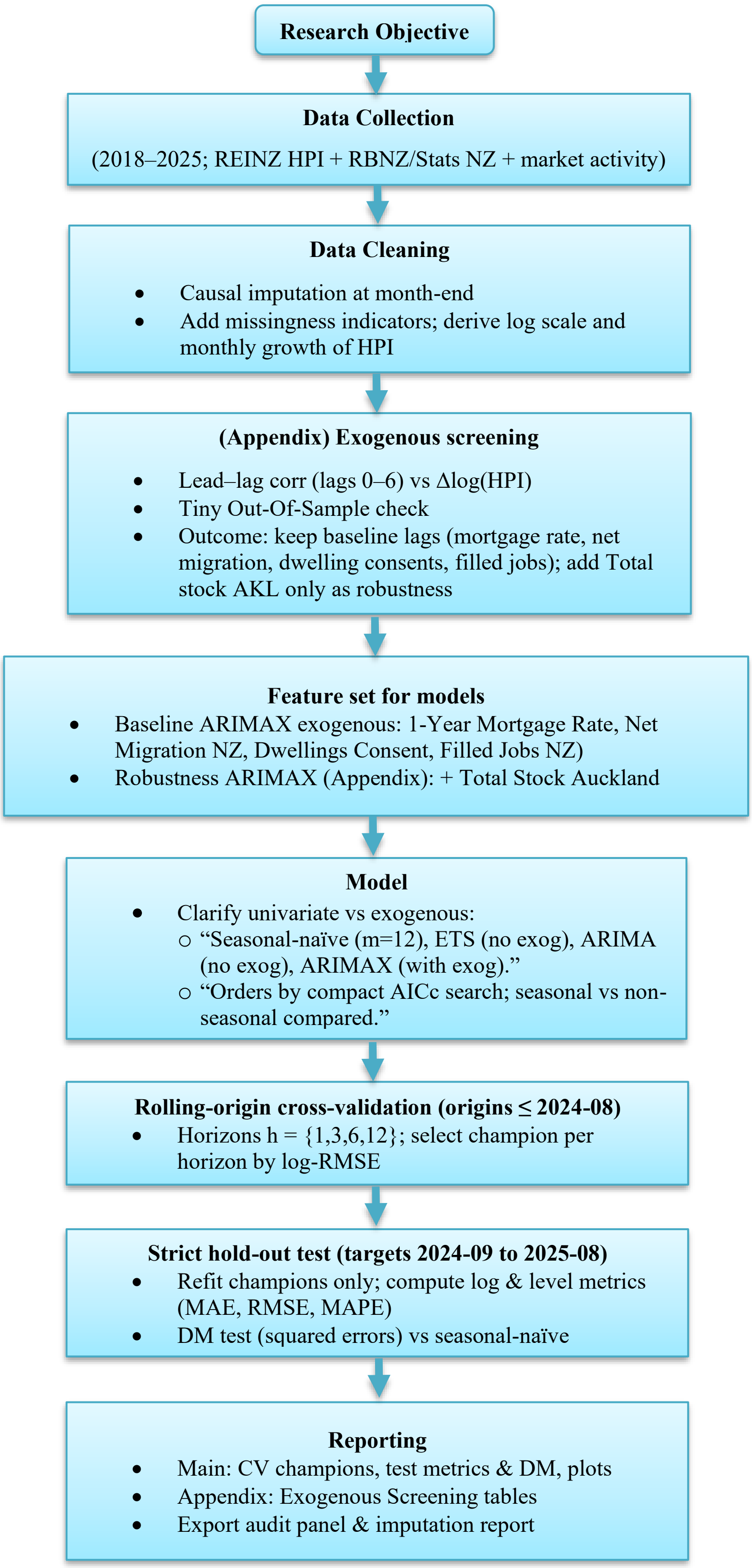
## 4. Research Methods and Design

The study's goal is to produce accurate and transparent short-horizon forecasts of Auckland's monthly HPI and to test whether a compact, causally lagged set of economically motivated drivers improves accuracy over univariate baselines. Monthly housing series are short and exhibit regime shifts, so the design prioritises leak-free data handling, generalisation across regimes, and interpretability. Concretely, the pipeline aligns all inputs to month-end, applies strictly causal imputation, selects models using rolling-origin cross-validation (CV), and conducts a final audit on a time-separated hold-out window. Pre-committing "champions" by horizon through CV prevents post-hoc model selection and maps directly to the research questions about forecast accuracy at h ∈ {1, 3, 6, 12} and the incremental value of exogenous information.

### 4.1 Research Design and Methodology

The following figure 4.1 summarises the end-to-end workflow from data assembly to evaluation. Methodologically, the comparative set contains a seasonal-naïve benchmark, exponential smoothing (ETS), univariate ARIMA/SARIMA, and ARIMAX/SARIMAX with lagged exogenous regressors. The seasonal-naïve forecast $\hat{y}_{\boldsymbol{t+h|t}} = \boldsymbol{y_{t+h-12}}$ provides a stringent, interpretable floor. ETS is estimated with additive components when the sample length supports stable seasonal extraction. ARIMA/SARIMA orders are chosen from a compact grid by AIC, which is robust for modest monthly samples, while ARIMAX/SARIMAX uses the same grid but augments the mean equation with causally lagged drivers to avoid leakage.

Strengths of this design include leak-safe feature timing and imputation, horizon-wise pre-commitment via CV, and use of transparent, auditable models; its chief limitations are the short test window—which reduces statistical power at longer horizons—and the fact that ARIMAX coefficients are predictive rather than causal, while an aggregate HPI can still conceal composition shifts. This balance of methods and safeguards is appropriate for a single regional monthly series subject to structural change. This model set is used since they are standard, auditable baselines in housing forecasting, perform well on short monthly series, and make leakage control straightforward.



**Research Objective**

**Data Collection**

(2018–2025; REINZ HPI + RBNZ/Stats NZ + market activity)

**Data Cleaning**

- Causal imputation at month-end
- Add missingness indicators; derive log scale and monthly growth of HPI

**(Appendix) Exogenous screening**

- Lead–lag corr (lags 0–6) vs Δlog(HPI)
- Tiny Out-Of-Sample check
- Outcome: keep baseline lags (mortgage rate, net migration, dwelling consents, filled jobs); add Total stock AKL only as robustness

**Feature set for models**

- Baseline ARIMAX exogenous: 1-Year Mortgage Rate, Net Migration NZ, Dwellings Consent, Filled Jobs NZ)
- Robustness ARIMAX (Appendix): + Total Stock Auckland

**Model**

- Clarify univariate vs exogenous:
  - “Seasonal-naïve (m=12), ETS (no exog), ARIMA (no exog), ARIMAX (with exog).”
  - “Orders by compact AICc search; seasonal vs non-seasonal compared.”

**Rolling-origin cross-validation (origins ≤ 2024-08)**

- Horizons h = {1,3,6,12}; select champion per horizon by log-RMSE

**Strict hold-out test (targets 2024-09 to 2025-08)**

- Refit champions only; compute log & level metrics (MAE, RMSE, MAPE)
- DM test (squared errors) vs seasonal-naïve

**Reporting**

- Main: CV champions, test metrics & DM, plots
- Appendix: Exogenous Screening tables
- Export audit panel & imputation report

Figure 4.1: The Flowchart of Research Design and Methodology

### 4.2 Data Collection

The target series is the REINZ Auckland House Price Index observed monthly from January 2018 to August 2025 and aligned to calendar month-end. The target enters models on the log scale, $\boldsymbol{y_t} = \mathbf{log}\,(\boldsymbol{HPI_t})$, which stabilises variance and allows approximate percentage interpretation of errors. The core exogenous regressors are selected for economic relevance and timeliness: the New Zealand one-year fixed mortgage rate captures credit conditions that shift demand and expectations; national net migration proxies population pressure; Auckland residential building consents represent the supply pipeline in a post-up-zoning environment; and filled jobs track labour-market momentum linked to income and confidence. A wider set of market-activity indicators—OCR, new listings, total stock of listings, sales count, days to sell, and the median price series—was also assembled for screening; these variables are often strongly contemporaneous with prices, so they are evaluated carefully for leakage and incremental value and are not part of the baseline unless earned through out-of-sample gains. Screening details and a small out-of-sample check are reported in Appendix A (Tables A1–A2; heatmaps). Data Sources are from the Reserve Bank of New Zealand, RBNZ Statistics (mortgage rates[1], OCR[2]), Stats NZ (net migration[3], building/ dwelling consents[4], filled jobs[5]), REINZ[6] (HPI, sales count, days-to-sell, median price), Realestate.co.nz[7] (new listings, total stock of listings).

[1] This is the link for *New residential mortgage interest rate* from Reserve Bank of New Zealand. https://www.rbnz.govt.nz/statistics/series/exchange-and-interest-rates/new-residential-mortgage-standard-interest-rates

[2] This is the link for *Official Cash Rate* from Reserve Bank of New Zealand. https://www.rbnz.govt.nz/monetary-policy/monetary-policy-decisions

[3] This is the link for *International net migration (July 2025)* from Stats NZ. https://www.stats.govt.nz/information-releases/international-migration-july-2025/

[4] This is the link for *Building/ dwelling consents (July 2025)* from Stats NZ. https://www.stats.govt.nz/information-releases/building-consents-issued-july-2025/

[5] This is the link for *Employment indicators: filled jobs (July 2025)* from Stats NZ. https://www.stats.govt.nz/information-releases/employment-indicators-july-2025/

[6] This is the link for *Property reports: Auckland HPI, sales, days to sell, and median price* from Real Estate Institute of New Zealand. https://www.reinz.co.nz/Web/Web/Data-and-Products/Property-reports.aspx

[7] This is the link for *Property Report: Auckland's new listings and total stock* from Realestate.co.nz. https://news.realestate.co.nz/blog/tag/the-new-zealand-property-report

### 4.3 Data Analysis Process

The analysis was conducted in Python using standard time-series libraries and proceeds through five linked stages designed to prevent leakage, enable fair model comparison, and keep the workflow auditable.

Step 1 is causal preprocessing and audit artefacts. All series are aligned to month-end before transformation. Missing values are imputed using only information available at the time: interest-rate series are forward-filled or interpolated within plausible bounds; activity counts (consents, listings, stock, sales, days-to-sell, median price) use a causal seasonal rule —same month last year, otherwise the historical month-of-year median, then forward-fill; filled jobs are forward-filled as a level series; net migration is forward-filled allowing negatives. The target is log(HPI) for variance stabilization and store *_imp and *_miss companions to preserve an audit trail.

Step 2 is lead–lag screening of additional indicators. To decide whether any of the market-activity variables should augment the baseline ARIMAX, each candidate is screened outside the main model. We compute lead–lag correlations between Δlog(HPI) and each market-activity variable across lags 0–6 and then run a one-variable out-of-sample check (Δlog(HPI) on each candidate's best causal lag) against a zero-change baseline. Strong contemporaneous co-movement (e.g., sales, new listings at lag 0) is treated as leakage and excluded. In the current window, new listings at lag 2 shows a small standalone gain, but this advantage does not persist once the variable is embedded in the full ARIMAX under cross-validation; therefore, screened indicators remain in the Appendix as robustness items rather than in the baseline model.

Step 3 is featuring construction for ARIMAX. The exogenous matrix comprises only leak-safe lags justified a priori—mortgage rate (L1), net migration (L2), dwelling consents (L4), and filled jobs (L1). No contemporaneous signals are used; optional lagged missingness flags can absorb data-quality effects.

Step 4 is model selection via rolling-origin cross-validation. Using expanding windows up to August 2024, each candidate model (Seasonal-naïve, ETS when feasible, ARIMA/SARIMA, and ARIMAX/SARIMAX with the compact exogenous set) generates forecasts at horizons h ∈ {1,3,6,12}. Selection is based on log-scale RMSE at each horizon; Diebold–Mariano tests versus the Seasonal-naïve to quantify differences in predictive accuracy.

Step 5 is strict hold-out audit. Horizon-specific champions pre-committed in cross-validation are re-estimated and evaluated on the time-separated test window (targets September 2024 to August 2025). We report MAE, RMSE, and MAPE on the log scale for comparability and on the level scale (HPI units) for interpretability, alongside DM tests versus the Seasonal-naïve. Plots of actual versus champion forecasts by horizon provide a visual check on turning-point behavior and error growth.

### 4.4 Research Process

The method was developed iteratively. An initial pilot established month-end alignment, causal imputation rules, and sanity plots. A compact ARIMA grid and an ETS baseline were then implemented and contrasted under non-seasonal and seasonal settings.

Screening of additional indicators followed, using the lead–lag and one-variable out-of-sample checks described above; based on these diagnostics and early CV runs, only the four pre-registered drivers were retained in the baseline ARIMAX, with screened variables documented as robustness in the Appendix. Rolling-origin CV produced horizon-specific champions, which were then audited on the strict hold-out window. Occasional supervisor consultations were used to confirm leakage handling, horizon definitions, and reporting artefacts, and to ensure the pipeline remained focused on the stated research questions.

### 4.5 Ethical Considerations

All inputs are public, aggregated series from REINZ, RBNZ, Stats NZ, and realestate.co.nz. No personal or commercially sensitive information is used. Data are cited to their providers, steps are logged, outputs are versioned, and the exact modelling panel and evaluation tables are exported so that results can be independently checked.

## 5. Findings, Analysis and Discussion

This section reports the empirical results, interprets them against the research questions, and discusses how they align with—or depart from—prior evidence.

### 5.1 Findings

**Findings: Cross-validation Results**

Using rolling-origin cross-validation (2018-01 to 2024-08), the model comparison identified clear horizon-specific "champions." A compact univariate ARIMA consistently minimized log-RMSE at the 1-, 3- and 6-month horizons, whereas ARIMAX edged ahead only at 12 months during the CV phase, shown in table 5.1. This pattern suggests that most short-run predictability in the training window is embedded in the HPI's own autocorrelation and seasonal structure, with limited incremental value from the exogenous set when selection is performed out-of-sample.

Table 5.1: All CV models for each horizon (2018-01 - 2024-08)

| h (Forcast horizon in months) | Model | Seasonal | $\text{RMSE}_{\text{Log}}$ |
|---|---|---|---|
| **1** | **Arima** | **FALSE** | **0.0156** |
| **1** | Arima | TRUE | 0.0191 |
| **1** | Arimax | FALSE | 0.0172 |
| **1** | Arimax | TRUE | 0.0199 |
| **1** | ETS | FALSE | 0.0173 |
| **1** | Seasonal-naïve | FALSE | 0.1320 |
| **3** | **Arima** | **FALSE** | **0.0358** |
| **3** | Arima | TRUE | 0.0501 |
| **3** | Arimax | FALSE | 0.0427 |
| **3** | Arimax | TRUE | 0.0612 |
| **3** | ETS | FALSE | 0.0435 |
| **3** | Seasonal-naïve | FALSE | 0.1423 |
| **6** | **Arima** | **FALSE** | **0.0648** |
| **6** | Arima | TRUE | 0.1048 |
| **6** | Arimax | FALSE | 0.0749 |
| **6** | Arimax | TRUE | 0.1297 |
| **6** | ETS | FALSE | 0.0907 |
| **6** | Seasonal-naïve | FALSE | 0.1439 |
| **12** | Arima | FALSE | 0.1063 |
| **12** | Arima | TRUE | 0.2187 |
| **12** | **Arimax** | **FALSE** | **0.0974** |
| **12** | Arimax | TRUE | 0.2473 |
| **12** | ETS | FALSE | 0.2103 |
| **12** | Seasonal-naïve | FALSE | 0.1326 |

**Findings: The Strict Hold-out Test Results**

The strict hold-out test is a time-separated audit used after cross-validation to verify generalisation and prevent post-hoc model tuning on the test period. We kept one full year (Sep 2024–Aug 2025) completely out of view while we built and chose the models. Only after choosing our 'winner' models using earlier data we unlock that year and test them once, exactly as they would run in real life. This 'strict hold-out' check shows whether the models generalise to new months and avoids any temptation (or accidental bias) to tweak models using the test period. This test (targets 2024-09 to 2025-08) refined that picture. After refitting only the pre-committed champions, ARIMA remained the most accurate at h = 1, 3 and 6 ($\text{RMSE}_{\text{log}} \approx 0.0108$, 0.0199, 0.0255 respectively), in table 5.2. When translated into levels, typical errors were modest—about 36 HPI points at one month, rising to 85 points at six months—with percentage errors in the 1–2% range, shown in table 5.3. At the one-year horizon, however, the seasonal-naïve benchmark proved more reliable than either ARIMA or ARIMAX ($\text{RMSE}_{\text{log}} \approx 0.0278$; $\text{MAPE}_{\text{level}} \approx 2.15\%$), shown in table 5.3. Diebold–Mariano tests against seasonal-naïve confirmed this hierarchy: ARIMA's improvement at h=1 was marginally significant at the 10% level (p≈0.094), while differences at h=3 and h=6 were statistically indistinguishable; by construction the naïve model is the winner at h=12, shown in table 5.4.

Table 5.2: All models on Log-scale metrics for each horizon (Test window: 2024-09 - 2025-08)

| **h (Forcast horizon in months)** | **Model** | **$\text{MAE}_{\text{Log}}$** | **$\text{RMSE}_{\text{Log}}$** | **$\text{MAPE}_{\text{Log}}$** |
|---|---|---|---|---|
| **1** | **Arima** | **0.0097** | **0.0108** | **0.1198** |
| **1** | Seasonal-naïve | 0.0145 | 0.0177 | 0.1789 |
| **3** | **Arima** | **0.0163** | **0.0199** | **0.2003** |
| **3** | Seasonal-naïve | 0.0163 | 0.0233 | 0.2380 |
| **6** | **Arima** | **0.0204** | **0.0255** | **0.2509** |
| **6** | Seasonal-naïve | 0.0279 | 0.0328 | 0.3435 |
| **12** | Arimax | 0.0714 | 0.0740 | 0.8795 |
| **12** | **Seasonal-naïve** | **0.0212** | **0.0278** | **0.2615** |

Table 5.3: All models' Level metrics in HPI units for each horizon (Test window: 2024-09 - 2025-08)

| **h (Forcast horizon in months)** | **Winner Model** | **MAE$_{level}$ (HPI units)** | **RMSE$_{level}$ (HPI units)** | **MAPE$_{level}$ (%)** |
|---|---|---|---|---|
| **1** | **Arima** | **32.56** | **36.13** | **0.97%** |
| **1** | Seasonal-naïve | 49.42 | 60.33 | 1.47% |
| **3** | **Arima** | **54.25** | **66.25** | **1.62%** |
| **3** | Seasonal-naïve | 65.33 | 79.02 | 1.96% |
| **6** | **Arima** | **67.96** | **85.07** | **2.04%** |
| **6** | Seasonal-naïve | 93.92 | 110.56 | 2.83% |
| **12** | Arimax | 248.73 | 259.38 | 7.42% |
| **12** | **Seasonal-naïve** | **71.58** | **94.29** | **2.15%** |

Table 5.4: Statistical comparison vs seasonal-naïve (log scale) for each horizon (Test window: 2024-09 - 2025-08)

| **h (Forcast horizon in months)** | **Winner Model** | **DM Stats Vs seasonal-naïve** | **p-value** |
|---|---|---|---|
| **1** | Arima | −1.676 | 0.094 |
| **3** | Arima | −0.539 | 0.590 |
| **6** | Arima | −1.046 | 0.295 |
| **12** | Seasonal-naïve | 0 | 1 |

The following test-window plots make these results intuitive. Figure 5.1 (h=1) shows the ARIMA path closely shadowing the actual HPI line and capturing month-to-month direction changes; the gap widens only briefly around local peaks. Figure 5.2 (h=3) preserves broad turning points but exhibits smoother swings than the realized series, a typical artifact of multi-step ARIMA forecasts. Figure 5.3 (h=6) shows increasing amplitude mismatch: the model tracks the cycle but over- and under-shoots as horizons extend. Figure 5.4 (h=12) illustrates why the seasonal-naïve wins at a year: repeating last year's level provides a better guide to the 12-month-ahead pattern than extrapolating a single fitted dynamic, especially through the gentle trend shifts observed in 2025.

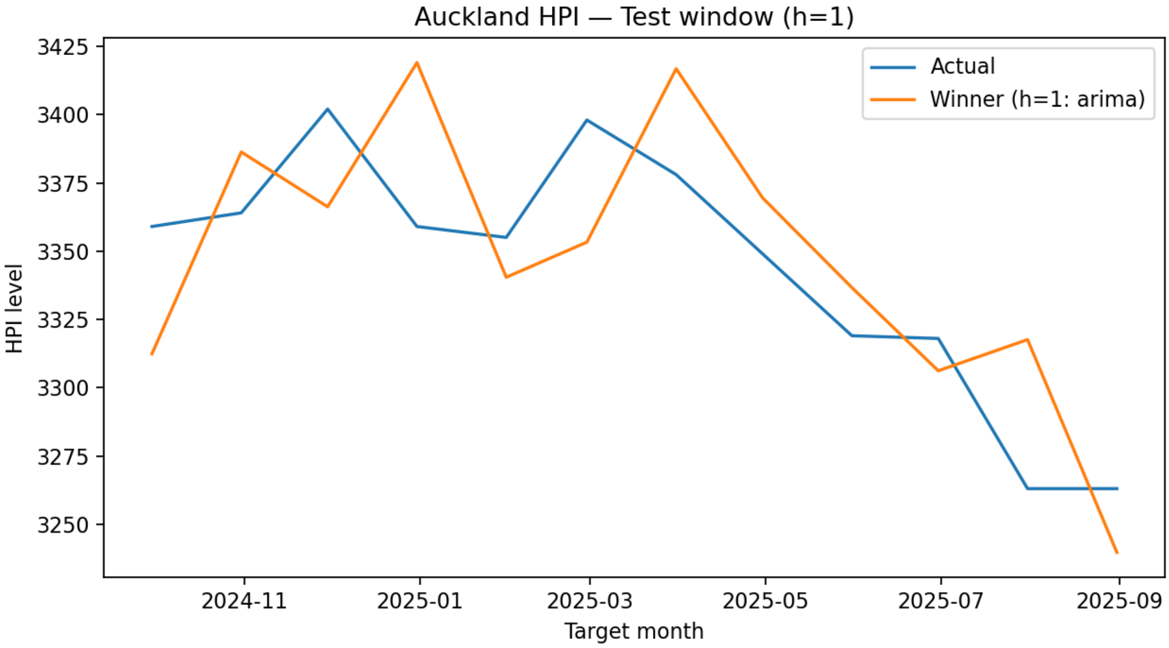


Figure 5.1: Auckland HPI test at h = 1 (2024-09 - 2025-08)

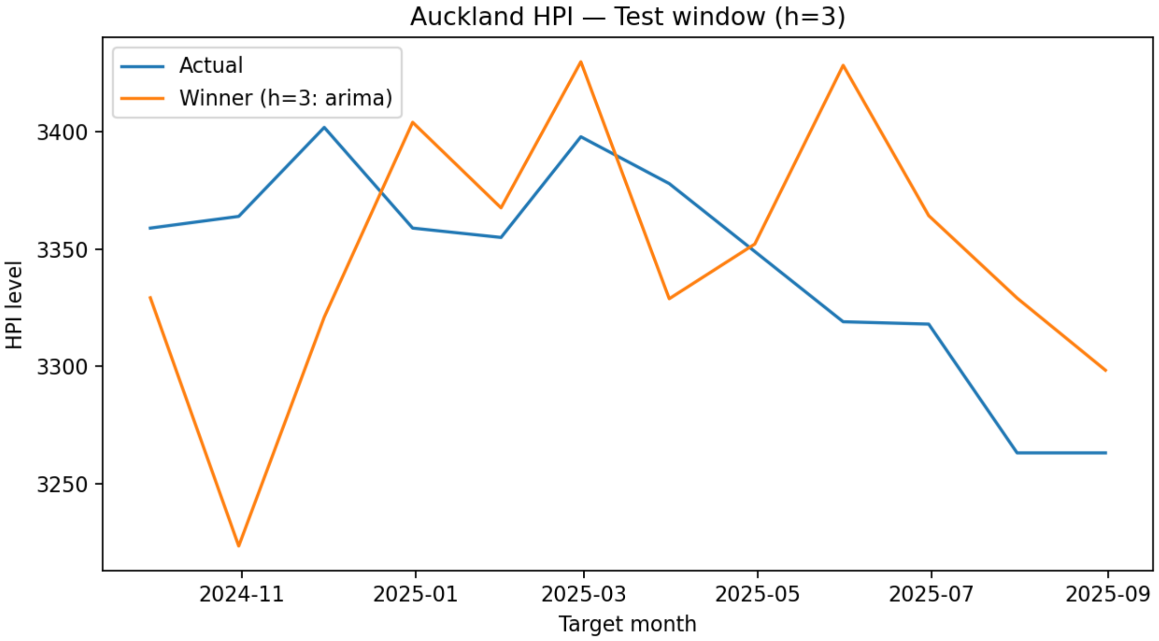


Figure 5.2: Auckland HPI test at h = 3 (2024-09 - 2025-08)

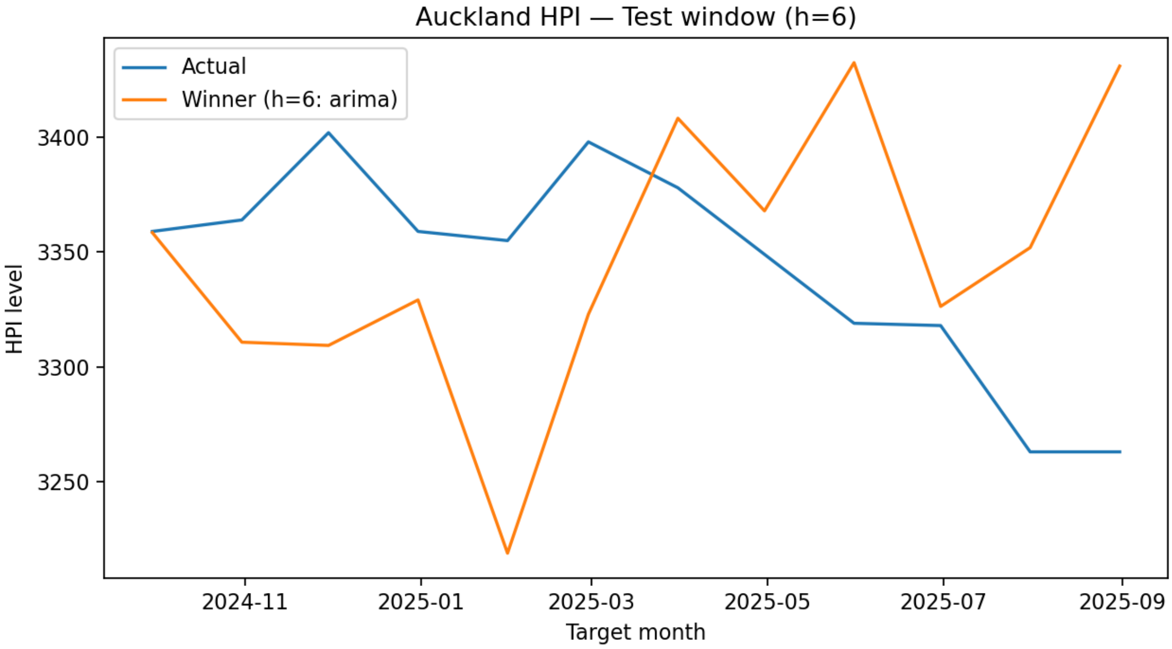


Figure 5.3: Auckland HPI test at h = 6 (2024-09 - 2025-08)

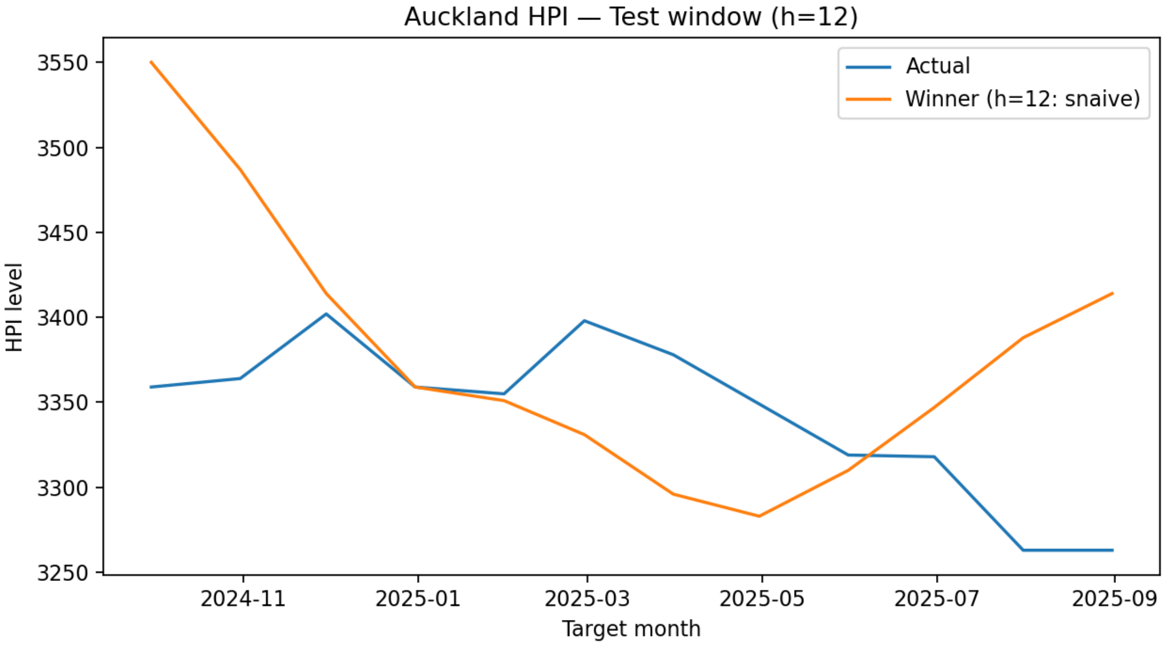


Figure 5.4: Auckland HPI test at h = 12 (2024-09 - 2025-08)

### 5.2 Analysis

When analyzing overall accuracy of the models forecast Auckland's monthly HPI at different horizons (main Research Question), transparent statistical baselines are sufficient for short-run HPI forecasts. Non-seasonal ARIMA dominates at 1–6 months with intuitive level-scale errors (≈1–2% MAPE), while seasonal-naïve is the most dependable one-year-ahead rule. This division of labor supports a pragmatic portfolio: ARIMA for operational nowcasting/quarter-ahead planning and seasonal-naïve for year-ahead anchoring unless updated evidence overturns the result.

When analyzing sub research question (a) for evaluating the value of exogenous drivers, we found that the compact, causally lagged driver set—mortgage rate (L1), net migration (L2), dwelling consents (L4), and filled jobs (L1)—did not yield robust out-of-sample gains over ARIMA on the test window. Although ARIMAX appeared favorable at h=12 in cross-validation, the advantage did not persist when audited on the disjoint from 2024-09 to 2025-08 targets, indicating seasonal-naïve beat at h=12 on the hold-out that any incremental signal was either unstable across regimes or eclipsed by the series' own seasonal persistence at that horizon.

When analyzing sub research question (b) for which horizons are most amenable to gains over the seasonal-naïve baseline, the clearest advantage over seasonal-naïve occurs at h=1, and even there significance is marginal, we found that improvements at h=3 and h=6 are present in point estimates but not statistically conclusive. At h=12, the annual seasonal repeat outperforms parametric dynamics in this period, consistent with many macro time-series where annual seasonality remains a powerful baseline.

When analyzing sub research question (c) we found no additional market-activity indicators (OCR, new listings, total stock, sales, days-to-sell, median price) worth including once screened for leakage and tested out-of-sample. Lead–lag heatmaps show strong contemporaneous co-movement for sales and new listings at lag 0, however, we exclude them to avoid leakage. A small one-variable out-of-sample check indicated that new listings at lag 2 reduced growth RMSE versus a zero-change benchmark, but this improvement did not persist when the variable was embedded in the cross-validated ARIMAX pipeline. Given potential endogeneity, multicollinearity with the pre-registered drivers, and the short 12-month test, none of the additional indicators is promoted to the baseline; they are retained only as appendix robustness candidates for future re-evaluation as the test period lengthens.

### 5.3 Discussion

These outcomes are consistent with prior research evidence that parsimonious Box–Jenkins models are formidable at short horizons in housing and related demand contexts, while simple seasonal rules remain hard to beat at longer spans in the presence of regime shifts (Alias, Zainun, & Abdul Rahman, 2016; Zainun, Mohamed Ghazali, & Mohd Sallehudin, 2016). In Auckland specifically, documented momentum from adaptive expectations and speculative behavior helps explain ARIMA's strong performance at $h \le 6$, yet the same mechanisms make distant turning points hard to predict, elevating seasonal-naïve at $h=12$ (Yang, Zhou, & Rehm, 2020). Contrary to expectation from work linking prices to credit/speculation and supply (Yang & Rehm, 2021; Greenaway-McGrevy & Phillips, 2023), our ARIMAX did not reliably outperform ARIMA on the hold-out. The most likely reasons are enforcing causal lags (reducing contemporaneous co-movement), structural change in 2024–2025, and aggregation/measurement frictions.

## 6. Conclusion

This study developed a transparent, reproducible pipeline for forecasting Auckland's monthly HPI that combines leak-safe data handling, rolling-origin cross-validation (from 2018-01 to 2024-08), and a strict target-month hold-out (from 2024-09 to 2025-08) by using seasonal-naïve, ETS, ARIMA, and ARIMAX. Across short horizons, univariate ARIMA performed best: at $h=1,3,6$ it achieved $\text{RMSE}_{\text{log}}$ of approximately 0.0108, 0.0199, 0.0255 with $\text{MAPE}_{\text{level}}$ around 0.97–2.04%; the improvement over the seasonal-naïve at $h=1$ was marginally significant ($p \approx 0.094$). At the annual horizon ($h=12$), however, the seasonal-naïve remained the most reliable benchmark ($\text{RMSE}_{\text{log}} \approx 0.0278$; $\text{MAPE}_{\text{level}} \approx 2.15\%$), indicating that simple seasonal persistence dominates one-year-ahead forecasts in this sample.

ARIMAX models with four causally lagged drivers did not deliver consistent gains on the hold-out, and additional screened indicators did not alter the champion set. The study contributes a horizon-specific accuracy map and a leak-aware forecasting template that others can replicate and extend.

## 7. Evaluation of the Research Method, Limitations & Future Work

The study sets out to (i) build a reproducible, leakage-safe forecasting pipeline for Auckland's HPI, (ii) benchmark transparent models across horizons, and (iii) test whether a compact set of exogenous variables adds value. The method delivered on those aims. Causal (forward-only) imputation and month-end alignment prevented look-ahead; rolling-origin cross-validation pre-committed a "champion" per horizon; and a time-separated hold-out provided an honest audit. Together these choices gave clear, comparable answers: ARIMA was most reliable at 1–6 months; the seasonal-naïve remained best at 12 months; and the compact ARIMAX did not consistently improve accuracy. The approach is also practical: the baselines are interpretable, easy to re-fit as new data arrives, and the code exports every table and figure required for replication. In short, the design matched the problem—short monthly series with possible regime shifts—by prioritising simplicity, auditability, and fair testing.

Three features were especially effective. First, the leakage controls (causal lags, no contemporaneous drivers) protected against overly optimistic results—a common risk in market activity data. Second, horizon-wise model selection avoided one-size-fits-all claims and produced actionable guidance for users who forecast at different lead times. Third, statistical comparison using Diebold–Mariano tests complemented the error metrics and helped interpret when the gains over the seasonal-naïve were meaningful versus merely numerical.

However, there are some limitations and their implication that it is worth noting. The 12-month hold-out limits statistical power at longer horizons, so some differences (especially at h=3 and h=6) are not decisive. The dependent variable is an aggregate HPI; while this reduces composition noise relative to medians, shifts in dwelling type or location can still blur monthly signals. Exogenous series (mortgage rates, migration, consents, filled jobs) are proxies for financing, demand, supply pipeline, and labor-market momentum; they are plausibly endogenous to prices, so coefficients are predictive rather than causal. Parameter stability is also uncertain because the sample spans structural breaks (COVID-19, credit cycles), and the monthly frequency limits the universe of usable predictors while exposing the panel to subsequent data revisions. Finally, the series begins post-AUP (2018), restricting direct identification of pre/post zoning effects, and linear ARIMA/ARIMAX may underfit non-linear or regime-dependent dynamics.

Future work should address these constraints in several directions. Extending the sample before 2016 and lengthening the hold-out would support explicit pre/post AUP analysis and more powerful evaluation at annual horizons, including regime-slice reporting that contrasts pandemic with post-pandemic performance. Disaggregation by sub-market (e.g., houses vs apartments, board areas) and integration of a hedonic layer would better control for composition effects while preserving an aggregate forecast. Methodologically, time-varying and regime-switching specifications—state-space ARIMAX with drifting coefficients or Markov-switching structures—could capture evolving relationships; simple forecast combinations and calibrated density forecasts would add robustness and probabilistic guidance. Incorporating higher-frequency leading indicators (listing flows, auction clearance rates, search intensity) may improve nowcasting, and policy evaluation could proceed via difference-in-differences or synthetic controls around zoning and LVR changes. As an interpretable machine-learning comparator, researchers can try regularised trees (e.g., XGBoost with lagged features) once a longer test window is available. These steps build directly on the findings and keep the emphasis on transparent, verifiable improvements.

## Appendix A. Exogenous screening

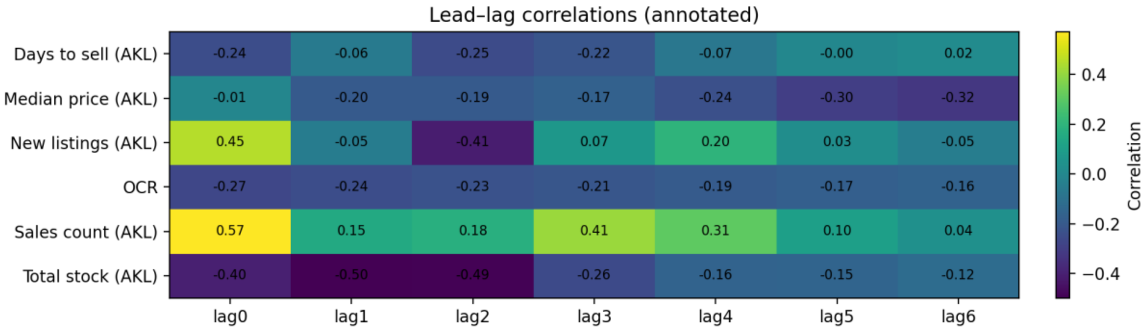


Figure A1. Lead-lag correlation heat map of exogeneous candidates and Δ log (HPI).

Table A1. Lead–lag correlations between candidate drivers and Δ log (HPI).
*(source: appendix_leadlag_corr.csv)*

| series | col | lag0 | lag1 | lag2 | lag3 | lag4 | lag5 | lag6 | best_lag | best_corr_signed | best_abs_corr |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **OCR** | ocr_imp | -0.27 | -0.24 | -0.23 | -0.21 | -0.19 | -0.17 | -0.16 | 0.00 | -0.27 | 0.27 |
| **New listings (AKL)** | new_listings_akl_imp | 0.45 | -0.05 | -0.41 | 0.07 | 0.20 | 0.03 | -0.05 | 0.00 | 0.45 | 0.45 |
| **Total stock (AKL)** | total_stock_akl_imp | -0.40 | -0.50 | -0.49 | -0.26 | -0.16 | -0.15 | -0.12 | 1.00 | -0.50 | 0.50 |
| **Sales count (AKL)** | sales_count_akl_imp | 0.57 | 0.15 | 0.18 | 0.41 | 0.31 | 0.10 | 0.04 | 0.00 | 0.57 | 0.57 |
| **Days to sell (AKL)** | days_to_sell_akl_imp | -0.24 | -0.06 | -0.25 | -0.22 | -0.07 | 0.00 | 0.02 | 2.00 | -0.25 | 0.25 |
| **Median price (AKL** | median_price_akl_imp | -0.01 | -0.20 | -0.19 | -0.17 | -0.24 | -0.30 | -0.32 | 6.00 | -0.32 | 0.32 |

Table A2. One-variable OOS RMSE of Δ log (HPI) on lagged drivers vs zero-change baseline.
*(source: appendix_oos_singlevar_rmse.csv)*

| series | col | chosen_causal_lag | RMSE_model | RMSE_naive0 | %_improvement_vs_naive0 | n_test_obs |
|---|---|---|---|---|---|---|
| **OCR** | ocr_imp | 1 | 0.01 | 0.01 | -12.16 | 12 |
| **New listings (AKL)** | new_listings_akl_imp | 2 | 0.01 | 0.01 | 23.26 | 12 |
| **Total stock (AKL)** | total_stock_akl_imp | 1 | 0.02 | 0.01 | -86.87 | 12 |
| **Sales count (AKL)** | sales_count_akl_imp | 3 | 0.01 | 0.01 | -9.26 | 12 |
| **Days to sell (AKL)** | days_to_sell_akl_imp | 2 | 0.01 | 0.01 | -5.17 | 12 |
| **Median price (AKL** | median_price_akl_imp | 6 | 0.01 | 0.01 | -6.61 | 12 |

Figure A1 and Table A1 show that the strongest associations with Δlog(HPI) occur contemporaneously (e.g., sales and new listings at lag 0), which we exclude to avoid leakage. Among causal lags, total stock at L1 exhibits the largest magnitude correlation (r≈−0.50), while new listings at L2 is moderate (r≈−0.41). However, the one-variable out-of-sample check in Table A2 indicates that total stock L1 worsens accuracy (≈−86.9% versus a zero-change growth baseline), and although new listings L2 reduces RMSE by ≈23.3%, this isolated gain does not survive in the multivariate ARIMAX under cross-validated selection or in the strict hold-out. Given the predominance of lag-0 co-movement, potential endogeneity/overlap with the pre-registered drivers, and the short 12-month test window, we do not promote any additional market-activity indicator. The baseline ARIMAX therefore retains mortgage rate L1, net migration L2, dwelling consents L4, and filled jobs L1; total stock L1 and new listings L2 are reported only as robustness variants (see Tables A1–A2).

## REFERENCES

Alias, A. R., Zainun, N. Y., & Abdul Rahman, I. (2016). Comparison between ARIMA and DES methods of forecasting population for housing demand in Johor. *MATEC Web of Conferences* (ICTTE 2016). https://doi.org/10.1051/matecconf/20168107002

Bade, D., Castillo, J. G., Fernandez, M. A., & Aguilar-Bohorquez, J. (2020). The price premium of heritage in the housing market: Evidence from Auckland, New Zealand. *Land Use Policy, 99*, 105042. https://doi.org/10.1016/j.landusepol.2020.105042

Cox, W. (2025). *Demographia international housing affordability: 2025 edition*. Chapman University, Center for Demographics and Policy. https://www.chapman.edu/communication/_files/Demographia-International-Housing-Affordability-2025-Edition.pdf

Dentons. (2024, July 5). *Government announces major plan for urban growth*. https://www.dentons.co.nz/en/insights/articles/2024/july/5/government-announces-major-plan-for-urban-growth

Greenaway-McGrevy, R., & Phillips, P. C. B. (2023). The impact of upzoning on housing construction in Auckland. *Journal of Urban Economics, 136*, 103555. https://doi.org/10.1016/j.jue.2023.103555

New Zealand Government—Beehive. (2025, February 28). *Going for Housing Growth: New and improved infrastructure funding and financing tools.* https://www.beehive.govt.nz/release/going-housing-growth-new-and-improved-infrastructure-funding-and-financing-tools

Nunna, K. C., Zhou, Z., & Shakya, S. R. (2023). Time series forecasting of U.S. housing price index using machine and deep learning techniques. *2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) (pp. 1–8). IEEE*. https://doi.org/10.1109/CSDE59766.2023.10487733

Rehm, M., & Yang, Y. (2021). Betting on capital gains: Housing speculation in Auckland, New Zealand. *International Journal of Housing Markets and Analysis, 14*(1), 72–96. https://doi.org/10.1108/IJHMA-02-2020-0010

Te Tūāpapa Kura Kāinga—Ministry of Housing and Urban Development. (2025, June 30). *Going for Housing Growth programme*. https://www.hud.govt.nz/our-work/going-for-housing-growth-programme

Te Tūāpapa Kura Kāinga—Ministry of Housing and Urban Development. (2025, Feb 17). *The Infrastructure Funding and Financing Act 2020*. https://www.hud.govt.nz/our-work/the-infrastructure-funding-and-financing-act-2020

Yang, Y., & Rehm, M. (2021). Housing prices and speculation dynamics: A study of Auckland housing market. *Journal of Property Research, 38*(4), 286–304. https://doi.org/10.1080/09599916.2021.1873405

Yang, Y., & Zhou, M. (2024). Estimating housing price bubbles for investment and owner-occupancy. *Applied Economics, 56*(55), 7339–7351. https://doi.org/10.1080/00036846.2023.2281292

Yang, Y., Zhou, M., & Rehm, M. (2020). Housing prices and expectations: A study of Auckland. *International Journal of Housing Markets and Analysis, 13*(4), 601–616. https://doi.org/10.1108/IJHMA-12-2019-0122

Zainun, N. Y., Mohamed Ghazali, F. E., & Mohd Sallehudin, M. S. (2016). Prediction of low-cost housing demand in Malaysia using ARIMA model. *MATEC Web of Conferences, 47*, 04008. https://doi.org/10.1051/matecconf/20164704008