Understanding the Housing Affordability Crisis: National Trends, Regional Disparities, and Economic Drivers in the U.S.

Jing Tan, Yanzhen Gao, Yanmin Gui, Jiayuan Gong

(DSAN-5100-01)

I. Abstract

Housing affordability has recently emerged as a growing policy concern in the United States, driven by returns in home prices, rents, and mortgage rates that are outpacing income growth for a substantial fraction of households. This paper undertakes a comprehensive statistical analysis of U.S. housing affordability on five thematic dimensions using monthly state-level data from 2000 to 2025: (1) structural change across the major economic cycles, (2) distributional stress measures, (3) uncertainty measures, (4) behavioral dynamic, and (5) fundamental economic drivers. The analysis examines the long-run shifts between homeownership and rental affordability, whether high-interest-rate regimes coincide with severe affordability crises; quantifies uncertainty using directed inference, models housing affordability dynamics using time-series methods, and examines spillovers between the rental and ownership housing markets. Research findings indicate that since 2013 and particularly since the outbreak of the pandemic, both home affordability and rental affordability have experienced significant structural deterioration, while home purchases have become more sensitive to macroeconomic shocks. Mortgage rates are strongly and reliably negatively related to housing affordability, and housing affordability shocks are temporally persistent, yet partly predictable. Overall, the findings present a multidimensional empirical perspective on how affordability has evolved over the last quarter century and provide important insights for policymakers and economists, and for households facing increasingly constrained housing markets.

II. Introduction & Background

Over the last quarter century, housing affordability has emerged as a key socioeconomic issue in the United States, as households from nearly every income segment have faced increasing financial pressures to secure stable housing. According to national housing groups, in most areas, home prices and rents have appreciated more quickly than wages of employment, which further exacerbates household financial stresses and heightens inequality in per-capita incomes (Joint Center for Housing Studies, 2023). While the cycles of housing affordability rise and fall with the economic cycle, the longer trends reveal continuing pressures on renters, potential homebuyers, and local communities. In fact, empirical studies confirm that mortgage rates, housing supply constraints, and income dynamics have weighed heavily on home affordability, though the mechanisms of change differ over time and between states (Freddie Mac, 2022). Together, these features make housing affordability a multilayered challenge that warrants careful empirical investigation.

There have been major structural shifts in the U.S. housing market since 2000, such as the house-price bubble of the first decade of the 21st century, the crisis of 2008, the low-interest recovery

of the 2010s, and more recent house-price growth with rising interest rates. It is established from previous literature that homeownership affordability is more sensitive to long-run house-price growth or mortgage rates, while rental affordability is more sensitive to shorter-term adjustment (Pew Research Center, 2021). However, few such research studies might have examined affordability gaps between such drastically different regimes or explored uncertainty in long-run affordability paths. Hence, the need for a more encompassing analysis for such research work.

To fill these research gaps, this research paper applies the methodology of testing the hypothesis, guiding inference, time series analysis, and regression analysis for the development of home affordability (HomeAff) and rental affordability (RentAff) using state-level data from 2000 to 2025. To transcend simple descriptive summaries, this study focuses on five core research directions:

1. **What structural shifts have occurred in homeownership versus rental affordability under major U.S. economic systems since 2000?**
2. **To what extent did periods of elevated mortgage rates coincide with severe affordability crises in the homeownership market?**
3. **When assessing uncertainty through bootstrapping, how robust are long-term affordability trends and key economic relationships?**
4. **What is the persistence of affordability shocks? Can time-series models predict short-term fluctuations?**
5. **How do rental and homeownership affordability fluctuate in tandem? To what extent are these patterns influenced by mortgage rates and real wages?**

This research explores five interlocking areas of study to better understand the underlying historical forces that influence the affordability of housing. It is through the understanding of interactions between macroeconomic factors, the cycle of mortgage rates, income trends, and the housing market that it hopes to provide pertinent findings for policy-makers, economists, planners, or any individual seeking to make their way through increasingly limited markets of housing purchase. This research hopes to not only understand trends of the past, but also their determinants that would influence the future of housing affordability in the USA.

III. Data

3.1 Research Design and Analytical Framework Overview

This study aims to systematically characterize the long-term evolution of U.S. housing affordability from 2000 to 2025. This paper constructs a hierarchical analytical framework centered on national-level monthly time series supplemented by state-level data. At the national level, the focus is on the long-term trends, cyclical variations, and relationship with interest rate

environments of homeownership affordability (HomeAff) and rental affordability (RentAff). At the state level, descriptive comparisons reveal regional heterogeneity potentially obscured by national averages.

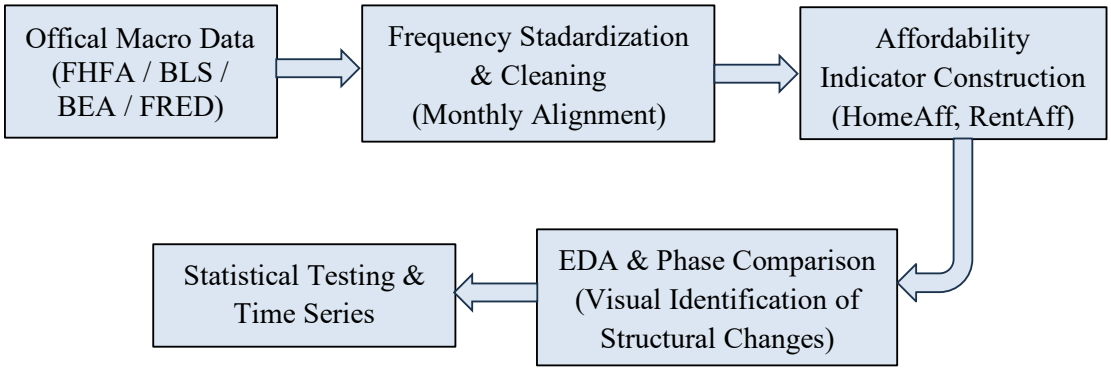


Table1: Process Design Diagram

3.2 Data Source

At the national level of macroeconomic data, this study is based entirely on publicly available secondary macroeconomic data and does not involve any original surveys. Core variables at the national level include housing prices, rents, income, and mortgage interest rates, sourced from official U.S. government or authoritative economic databases:

Source	Data	Variable name	Frequency	Description
FHFA	Purchase-Only House Price Index	$house\_price_t$	Monthly	This index is constructed using the repeat sales method and effectively reflects the cumulative trend in transaction prices for owner-occupied housing.
BLS	Rent of Primary Residence	$rent\_cost_t$	Monthly	Depict the actual cash rent levels paid by urban residents.
BEA	Disposable Persional Income pre capita	$wage\_income_t$	Monthly annualized	This indicator is published on an annualized basis and subsequently converted into monthly income flows.
FRED	30-Year Fixed Rate Mortgage Average in the United States		Weekly	The raw weekly financing environment data is converted into monthly interest rates through a simple monthly average.

Table2: Independent variable

Given that national averages may obscure regional disparities, this paper further incorporates the FHFA's quarterly state-level housing price index and BEA's state-level per capita disposable

income data. States with markedly different economic structures—such as D.C., Utah, Hawaii, Alabama, Mississippi, and West Virginia—are selected as examples. The state-level analysis serves solely for descriptive comparison, constructing an indicator of housing price growth relative to income growth to illustrate the heterogeneity in housing affordability trajectories across regions. Given limitations in the coverage and comparability of long-term state-level rental data, the state-level analysis does not separately construct a rental affordability metric nor incorporate it into formal statistical modeling.

3.3 Housing Affordability Index (dependent variable)

After completing frequency alignment, this paper constructs affordability indicators for homeownership and renting. Homeownership affordability is defined as the ratio of the house price index to income:

$$HomeAff_t = \frac{house\_price_t}{wage\_income_t},$$

Rental affordability is defined as the ratio of the rent index to income:

$$RentAff_t = \frac{rent\_cost_t}{wage\_income_t},$$

Under this definition, a higher indicator value indicates that housing or rental costs are relatively higher relative to income, meaning residents face a heavier burden.

In the quarterly state-level data, this study constructs two types of indicators to characterize regional relative affordability. The first type measures indicators relative to the national average for the current period. Let $HPI_{\{s,t\}}$ denote the purchase-only data HPI for state s in period t, $HPI_{\{US,t\}}$ notes the national purchase-only data HPI, $income_{\{s,t\}}$ and $income_{\{US,t\}}$ denote the income indicators for state s and the nation, respectively. Then, this study defines:

$$RelPrice_{\{s,t\}} = \frac{HPI_{\{s,t\}}}{HPI_{\{US,t\}}}, RelIncome_{\{s,t\}} = \frac{income_{\{s,t\}}}{income\{US,t\}}, AffordGap_{\{s,t\}} = \frac{RelPrice_{\{s,t\}}}{RelIncome_{\{s,t\}}}.$$

When $AffordGap_{\{s,t\}} > 1$, it indicates that within the state, the relative increase in housing prices compared to the national average has outpaced the relative increase in income compared to the national average.

The second category consists of indicators relative to the state's historical base period. Selecting the first quarter of 2000 as the base period, the housing price and income indices are defined as follows:

$$Price\ IndexSelf_{\{s,t\}} = \frac{HPI_{\{s,t\}}}{HPI_{\{S,t_0\}}}, Income\ IndexSelf_{\{s,t\}} = \frac{income_{\{s,t\}}}{income_{\{s,t_0\}}},$$

And further define:

$$AffordLevel_{\{s,t\}} = \frac{PriceIndexSepf_{\{s,t\}}}{IncomeIndexSelf_{\{s,t\}}}.$$

This indicator reflects the deterioration or improvement in home affordability within the state relative to the beginning of 2000. State-level analysis primarily employs descriptive comparisons based on the temporal trends of these indicators, without engaging in formal model estimation.

3.4 Macro-Period Classification and Categorical Variables

To compare systemic differences in housing affordability across distinct economic environments, this study divides the period from 2000 to 2025 into four macro-phases:

(1) the real estate bubble phase (2000–2007);
(2) the financial crisis and recovery phase (2008–2013);
(3) the low-interest-rate expansion phase (2014–2019);
(4) The post-pandemic and high-interest-rate period (2020-2025).

Accordingly, the categorical variable period is constructed for subsequent grouping descriptions, mean comparisons, and variance analysis.

Additionally, a binary indicator variable for "high-burden months" is constructed based on the sample median. This variable is used to analyze whether high-interest-rate environments are more likely to be accompanied by severe housing affordability challenges.

3.5 Data Cleaning and Preprocessing

Due to variations in frequency and measurement methods among raw datasets, data preprocessing prioritizes standardizing frequency as its core objective. National-level housing prices and rents are inherently monthly data; annualized disposable income is converted to monthly income flows by dividing by 12; weekly mortgage rates are averaged monthly. State-level quarterly data undergoes harmonization with income data while maintaining quarterly frequency.

Regarding missing values, a very small number of missing months resulting from release delays were directly excluded without imputation to avoid introducing additional economic assumptions. Extreme volatility observed during the 2008–2009 financial crisis and the 2020–2021 pandemic period is treated as part of genuine economic shocks. No truncation or tailing adjustments were applied, though this is explicitly noted in result interpretations.

During time series modeling, HomeAff underwent differencing based on stationarity test results to suppress trend components and improve residual properties. Specific transformation methods were determined jointly by statistical tests and model diagnostics.

3.6 Exploratory Data Analysis

This paper employs time series plots of HomeAff and RentAff constructed to visually identify long-term trends, cyclical fluctuations, and potential structural breaks. Appendix I: Figure 1

reveals significant structural shifts and a clear divergence between the two markets. While the Rent Affordability remains relatively stable with moderate variance, the Home Affordability Index exhibits extreme volatility, particularly during the 2008 crisis and the post-2020 period. While these patterns strongly suggest that the data is non-stationary and subject to regime switching.

Then, box plots grouped by macroeconomic phases compare affordability distributions across different periods.As illustrated in Appendix I: Figure 2, the boxplots show a clear upward structural shift in the current period. The current "Post-COVID" Home Affordability is significantly different from every prior period, validating that the housing market has entered a distinctly more expensive regime. The visual story in Appendix I: Figure 3 diverges from the housing market. While distinct regimes exist, the boxplots reveal a period of stability between the Recovery and Expansion phases. Furthermore, the structural break in the current "post-COVID" period was negative compared to the Expansion period.

State-level data are presented as line charts juxtaposed with the national average trajectory, illustrating regional deviations. According to Appendix I: Figure 3, across all six states and regions, housing affordability closely follows the national trend, with major turning points aligned during the 2008 financial crisis and the post-2020 surge. This synchronicity confirms that nationwide macroeconomic forces, particularly mortgage rates and monetary policy, govern the overall direction of housing cycles. However, substantial differences emerge in magnitude. D.C. and Hawaii exhibit a pronounced decoupling between housing prices and income, with Utah rising sharply above the national average after 2015, indicating amplified affordability stress in high-growth regions. In contrast, Mississippi and West Virginia remain near the unity line, where income growth has broadly kept pace with housing appreciation. Overall, the results indicate that national-level analysis provides a conservative benchmark and tends to understate the affordability risks faced by high-growth markets.

IV. Methodology

Question 1: Changes in Housing Affordability Over Time

Long-term changes in housing affordability are often closely linked to macroeconomic cycles. This study first examines whether housing affordability has significantly deteriorated during specific periods after removing short-term fluctuations, particularly whether it has surpassed levels seen during previous real estate bubbles in the post-pandemic phase.

To investigate whether home affordability for ordinary households in the post-2013 period was significantly worse than during the pre-crisis period from 2000 to 2007, this study employs a two-sample t-test to compare the average home affordability across the two periods. Prior to testing, samples undergo normality and homogeneity of variance checks; if assumptions are not met, the nonparametric Wilcoxon signed-rank test serves as an alternative.

At the same time, to investigate whether rental affordability (RentAff) significantly deteriorated in the post-COVID era (2020–2025) compared to the pre-COVID era (2013–2019), this study employs a two-sample t-test supplemented by normality and homogeneity of variance tests, with nonparametric alternatives used when necessary.

Question 2: The Relationship Between Phases and Housing Affordability

To explore whether different macroeconomic phases constitute markedly distinct housing affordability conditions, this study employs one-way ANOVA to compare mean differences across the four phases. Upon establishing overall significance, Tukey's post-hoc test identifies specific sources of variation. Should assumptions of normality or homogeneity of variance be significantly violated, the Kruskal–Wallis test serves as a nonparametric alternative.

Question 3: The Causal Relationship Between High Interest Rates and Housing Affordability

To test whether high-interest rate environments are more likely to be accompanied by severe housing affordability burdens, this study constructs binary classification variables for “high/low interest rates” and “high/low housing affordability” based on the sample median. A contingency table is formed, and statistical inference is conducted using the chi-square independence test. Since this issue focuses on whether high/low interest rates and high/low burdens exhibit a structural co-occurrence relationship rather than a linear relationship between continuous variables, the chi-square independence test is employed instead of correlation analysis. Hypotheses Setting:

H0	**Months with high interest rates are independent of months with high housing affordability burdens**
H1	**Months with high interest rates are statistically associated with months with high housing affordability burdens**

Question 4: Robustness of Housing Affordability Conclusions to Distribution Assumptions

To calculate the bootstrap confidence intervals for changes in average housing affordability from the early 21st century to the 2020s, this study employed the Bootstrap method to estimate the sampling distribution of mean differences. It constructed 95% confidence intervals while reporting standard errors and deviation estimates. This method avoids relying on the assumption of normality, providing more robust interval estimation for mean differences when macroeconomic time series exhibit fat tails and extreme values.

To assess the strength of the correlation between mortgage interest rates and housing affordability throughout the entire sample period, this study employs Bootstrap resampling on paired observations of interest rates and HomeAff to construct the empirical distribution of the correlation coefficient along with its 95% confidence interval. By resampling observation pairs,

the Bootstrap method can assess the sensitivity of correlation conclusions to outliers and distribution assumptions.

Question 5: Dynamic Characteristics and Predictability of Housing Affordability Shocks

To explore this issue in greater detail, this study has broken it down into three sub-questions.

First, how persistent are shocks to home purchase affordability, and can we forecast near-term changes? This study first assesses stationarity using time series plots, ACF, PACF, and ADF unit root tests; differencing is applied when necessary. Subsequently, ARIMA or seasonal ARIMA models are constructed to model the dynamic structure of HomeAff. Stationarity was assessed using the Augmented Dickey–Fuller test, where failure to reject the null indicates the presence of a unit root. Model adequacy is assessed through residual plots, residual ACF, and Ljung–Box tests. Upon successful model validation, short-term forecasts spanning 12–24 months are generated. If the null hypothesis of residual independence cannot be rejected, it is considered that the model has effectively extracted the systematic time structure. Hypothesis setting:

H0 (Unit Root Test)	**The HomeAff series exhibits a unit root (non-stationary)**
H1	**The HomeAff sequence is stationary or can be made stationary through differencing**

Second, how synchronized are changes in home and rent affordability over time? By analyzing the synchrony between HomeAff and RentAff through time-series parallel plots and cross-correlation functions (CCF), and comparing fluctuations in amplitude across different phases, this study assesses whether the rental market serves as a buffer when homeownership affordability deteriorates.

Third, to what extent can changes in home affordability be explained by mortgage rates and real wages? This study constructs an exploratory regression model, regressing HomeAff on mortgage interest rates and income variables, employing different forms when necessary to mitigate non-stationarity issues. This analysis serves solely to describe correlations and does not draw causal inferences.

V. Result

5.1 Question 1: Changes in Housing Affordability Over Time

To test whether the change in the U.S. real estate market is a structural change or a cyclical fluctuation, this analysis first looks at the historical affordability index. There are obvious fluctuations in Appendix I: Figure 1, and hypothesis testing needs to be used to quantify whether the current economic environment has significantly deteriorated compared with the historical

benchmark. Assume that the current crisis is unique, and there are statistical differences between the average housing affordability index after 2020 and the pre - bubble level in 2008.

5.1.1 Homeownership Affordability in Early Bubble and the Recent Crisis

This study compares the early period of the housing bubble from 2000 to 2007 with the more recent period after 2020 to test the hypothesis of institutional changes in the cost of ownership. A Welch's two-sample t-test yielded a p-value of 1.201e-15, well below the 0.05 significance level. This result strongly rejects the null hypothesis, confirming a statistically significant difference between the two periods.

Importantly, the sample estimation analysis reveals a significant structural increase: the average affordability index increased from 0.435 in the early period to 0.527 in the more recent period. From a market structure perspective, this confirms that the current crisis is not a temporary fluctuation but a shift towards a structurally higher cost level, imposing a significantly heavier economic burden on households than the baseline levels during the historical housing bubble period.

5.1.2 Rental Affordability in Expansion and Recent Crisis

A parallel analysis was conducted on the rental market to investigate whether rental burdens have also increased. The Rental Affordability Index during the expansion period from 2013 to 2019 was compared to the recent period from 2020 to 2025. A Welch's two-sample t-test yielded a t-statistic of -13.631 with a p-value < 2.2e-16, indicating a highly significant difference. Sample estimates show that the average rental index increased from 0.862 in the earlier period to 1.033 in the recent period. These results are not in line with the hypothesis of a stable rental market, showing a serious structural decline in the rental sector along with the development of the property ownership market.

5.2 Question 2: The Relationship Between Phases and Housing Affordability

To further illustrate the evolution of this structural shift, we analyzed whether four defined macroeconomic phases (bubble, recovery, expansion, and post-pandemic era) represent statistically distinct stages. This analysis employed one-way analysis of variance (ANOVA) to test the stability across the entire timeline. The model yielded an F-value of 316.2 and a p-value less than 2e-16, confirming that the differences between different economic periods are significantly greater than the differences within each period.

The Tukey HSD test confirmed that housing affordability in the current post-pandemic era is significantly different from each of the previous periods, validating that the housing market has entered a significantly higher price phase.

5.3 Question 3: The Causal Relationship Between High Interest Rates and Housing Affordability

This section goes beyond current trends by examining distributional pressures through an analysis of whether a high-interest-rate environment could structurally trigger a housing affordability crisis. This analysis uses the Pearson chi-squared test to examine the relationship between a high-interest-rate environment and high housing unaffordability. The test results show a chi-squared statistic of 1.0658 and a p-value of 0.3019.

Despite the statistical result, this relationship is vividly depicted in the heatmap in Appendix I: Figure 5. The darker shading within the quadrant corresponding to high mortgage rates and high home affordability indices highlights the strong dependency. This suggests that while the relationship may not be deterministic, high-rate regimes disproportionately increase the probability of the market entering a crisis zone. This implies a non-linear threshold effect: once mortgage rates cross a certain level, the market structurally shifts into a high-burden category, acting as a primary driver of the structural break observed in 2020.

5.4 Question 4: Robustness of Housing Affordability Conclusions to Distribution Assumptions

To assess the reliability of these structural and distributional findings, this section applies Bootstrap inference. Given that economic time series often exhibit fat tails and non-normal distributions which can skew standard parametric tests, bootstrapping provides a method to quantify the uncertainty surrounding the observed shifts and correlations. The structural break identified in Question 1 was validated by bootstrapping the difference in means between the early 2000s and the 2020s as shown in Appendix I: Figure 6.

Appendix I: Figure 6 displays the resulting distribution of these 5,000 mean differences. The histogram is clearly centered away from zero, with the entire probability density mass shifted to the positive side. The 95% confidence interval derived from the quantiles strictly excludes zero, providing robust, distribution-free confirmation that the recent deterioration in affordability is a genuine structural shift rather than a statistical artifact of specific outliers.

Similarly, the correlation between mortgage rates and affordability was bootstrapped in Appendix I: Figure 8. The distribution of correlation coefficients is tightly clustered and far removed from zero. This confirms that the negative relationship between rates and affordability is a stable economic law, robust across different subsamples of the data, and reinforces the validity of the distributional stress observed in the previous section.

5.5 Question 5: Dynamic Characteristics and Predictability of Housing Affordability Shocks

Based on the structural break analysis above, which identified significant regime shifts in the post-2020 era, modeling of the future dynamics of housing affordability proceeded. Given the

extreme volatility identified in the housing market, this section prioritizes a rigorous examination of the Home Affordability Index to capture its complex temporal behavior. Subsequently, the same methodology is applied to the Rent Affordability Index to serve as a comparative baseline, quantifying the future divergence between the ownership and rental markets.

5.5.1 Model Identification: Decomposition and Stationarity

To distinguish systemic dynamic structures from short-term noise, the analysis first uses a classic multiplicative decomposition method. Appendix I: Figure 8 isolates the long-term trend component, which clearly shows the evolution of market structure. Importantly, the extracted trend is not linear but aligns with our phase divisions, showing distinct inflection points around 2008, 2014, and 2020. This confirms that the underlying "affordability base" undergoes structural changes with major macroeconomic cycles, reflecting transitions between bubble, recovery, expansion, and post-pandemic eras.

Combining trend analysis, Appendix I: Figure 9 displays the extracted seasonal component. The temporal trend appears to show a repeated pattern with a fixed yearly period. The frequent fluctuations suggest that housing affordability is not only governed by random economic shocks but also by predictable intra-year cycles. The twelve-month period indicates the modeling: for this repeated dynamic behavior, a seasonal difference (D=1) must be introduced in SARIMA to stabilize the mean.

Finally, Appendix I: Figure 10 displays the residual component, representing the stochastic noise left after extracting trend and seasonality. A critical insight from this plot is the **volatility clustering** observed post-2020, which coincides with the COVID-19 pandemic. This marked increase in residual variance indicates that the recent era is characterized by profound economic shocks and heightened uncertainty, suggesting that the market's dynamic behavior has become less predictable compared to the stable expansion phase.

5.5.2 Correlation Structure and Stationarity Check

To further explore the internal memory and persistence of the market, Lag plots and an Autocorrelation Function (ACF) plot were generated. Appendix I: Figure 11 displays a strong positive linear relationship at Lag 1 and Lag 2, indicating high persistence—meaning current affordability is heavily dependent on its immediate past. The ACF plot in Appendix I: Figure 12 corroborates this with a slow, linear decay of correlation coefficients. In the context of market dynamics, this pattern is a hallmark of a non-stationary process, implying that shocks to home affordability are permanent rather than mean-reverting; once affordability deteriorates, it does not naturally bounce back to a historical average without external correction.

This visual evidence of persistence was formally verified using the Augmented Dickey-Fuller test. The test yielded a p-value of 0.91, failing to reject the null hypothesis of non-stationarity. Consequently, differencing was required to stabilize the series, transforming the modeling focus from absolute levels to rates of change before SARIMA modeling could proceed.

5.5.3 Differencing

To address the identified non-stationarity and isolate the short-term dynamic structure, first-order differencing was applied. As shown in Appendix I: Figure 13, this transformation effectively removes the stochastic trend, resulting in a series that fluctuates around a constant mean of zero. A subsequent ADF test yielded a p-value of 0.01, confirming that stationarity had been achieved. This step is critical, as it ensures that the subsequent forecasting model captures the intrinsic dynamics of affordability changes rather than spurious correlations driven by historical trends.

5.5.4 Model Identification and Selection

With the series stationary, the ACF and PACF plots of the differenced data were analyzed to identify the specific autoregressive and moving average structures governing the market. Appendix I: Figure 14 reveals significant spikes at Lag 1 and Lag 12. The significance at lags 0 and 1 indicates a non-seasonal moving average component (q=0 or 1), capturing immediate shock adjustments. Crucially, the distinct spike at Lag 12 serves as a canonical signature of a seasonal moving average process (Q=1). This confirms that the market's dynamic behavior is characterized by a seasonal memory, where conditions from exactly one year prior significantly influence current changes.

The PACF plot in Appendix I: Figure 15 further refines this identification, showing spikes that suggest candidate autoregressive orders. Based on these temporal signatures, several candidate models were fitted.

Table 3 presents the comparison based on AIC and BIC values. The selection process prioritized parsimony: the comparison indicates that SARIMA (0,1,0) (0,1,1) [12] provides the optimal balance between model complexity and goodness-of-fit, best capturing the underlying data generation process.

Model	AIC	BIC
SARIMA(0,1,1)(0,1,1)[12]	-3039.502	-3024.754
SARIMA(1,1,0)(1,0,0)[12]	-3053.933	-3039.185
SARIMA(1,1,1)(0,1,1)[12]	-3054.909	-3036.475
SARIMA(0,1,0)(0,1,1)[12]	-3008.185	-2997.124

Table 3: Manual Model Comparison

5.5.5 Diagnostic Validation

To rigorously validate the manual model selection, it was benchmarked against an automated algorithm called auto.arima, which proposed a significantly more complex specification. Appendix I: Figure 16 displays the diagnostics for the Automated Model, where high complexity raised concerns about overfitting—modeling the noise rather than the signal.

In contrast, Appendix I: Figure 17 presents the diagnostics for the Manual Model. Visually, the residual behavior is nearly identical to the automated model, demonstrating that the simpler structure is equally capable of extracting the temporal dynamics. Quantitatively, the manual model achieves a lower AIC of approximately -10.18 compared to the automated model's -10.17. Given that the manual model achieves a superior fit statistic with fewer parameters, it was identified as the optimal choice.

Focusing on the manual model selected in Appendix I: Figure 17, the diagnostic panel at the top confirms its statistical sufficiency, just like the standardized residuals shown in the figure. They do not show a particularly prominent trend or system pattern, although these residuals reflect a situation where fluctuations have intensified after 2020. The residual autocorrelation function displayed in the left middle panel can confirm that the time dependence has been effectively eliminated. This is because all the lag cases fall within the 95% confidence interval, which indicates that the residuals exhibit white noise behavior. The normal Q-Q graph displayed in the right middle panel shows that the residuals generally conform to the theoretical normal distribution. However, due to the impact of extreme market shocks, the residuals only deviate slightly at the tail end. A very crucial point is that the P-value of the Ljung-Box test has always been higher than the maximum threshold, which confirms that we cannot reject the null hypothesis of independence.

5.5.6 Benchmarking Performance

To assess accuracy, the chosen SARIMA model was tested against a Seasonal Naïve benchmark on a held-out test set. Appendix I: Figure 18 shows the forecast performance. Unlike the naive benchmark, the SARIMA model tracks the recent decline and volatility with greater accuracy, capturing the market's turning points. Quantitative metrics confirm this superiority, with lower RMSE and MAE. This shows that despite high volatility; the housing market still has a stable dynamic structure and can be leveraged for short-term forecasts.

5.5.7 Final Forecast and Interpretation

Finally, the model was refitted on the full dataset to generate a 6-month forecast. Appendix I: Figure 19 projects the index with 95% and 99% confidence intervals. The forecast indicates that home affordability will remain strained and highly volatile in the near term. Most critically, the fan-shaped widening of the confidence intervals reflects the intrinsic uncertainty identified in the earlier decomposition. This implies that the market is currently in a highly sensitive state where new macroeconomic shocks could lead to significant deviations, suggesting that the structural pressure on homebuyers is unlikely to resolve immediately or linearly.

5.5.8 Comparative Analysis: Rent Affordability

Parallel to the housing market, the same modeling pipeline was applied to the Rent Affordability Index to highlight structural differences.

The decomposition of the rent index in Appendix I: Figure 20 reveals a significantly fluctuating trend compared to the housing market, though seasonality remains evident in Appendix I: Figure 21. Like the home index, the raw rent series was non-stationary as shown in Appendix I: Figure 22 and required first-order differencing to achieve stationarity, as confirmed by a subsequent ADF test.

Following the identification of model orders from the ACF/PACF plots shown in Appendix I: Figure 23, 24, the optimal SARIMA model SARIMA(1,1,1)(0,1,1)[12] was fitted. Diagnostic checks visualized in Appendix I: Figure 24 confirmed the validity of the model, with Ljung-Box test p-values consistently exceeding 0.05, indicating independent residuals.

The forecasting results reveal a stark contrast in market predictability across the two sectors. While the Home Affordability Index shown in Appendix I: Figure 19 displays a widening forecast interval indicative of high uncertainty, the Rent Affordability Index in Appendix I: Figure 25 follows a stable, mean-reverting path with relatively narrow confidence bands.

This quantitative difference implies a fundamental divergence in market dynamics: key drivers of rental prices appear to have stabilized, whereas homeownership affordability remains highly sensitive to external shocks. Taken together, these findings support the conclusion that the relationship between buying and renting dynamics has weakened, with the rental market currently serving as a stable buffer against the extreme volatility observed in the ownership market.

5.5.9 Joint Dynamics and Lead–Lag Patterns in Housing Affordability

This section studies the dynamic relationship between home and rent affordability, with a focus on differences in volatility, timing, and adjustment speed across the two markets. The analysis begins with a visual comparison of standardized affordability indexes to establish broad co-movement and regime shifts. It then proceeds to formal time-series tests and lag-based analyses to evaluate whether changes in one market systematically precede changes in the other. Together, these steps provide a structured framework for understanding how affordability evolves jointly over time.

5.5.10 Standardized Joint Dynamics of Home and Rent Affordability

This section examines how home and rent affordability evolve over time and whether one series systematically moves earlier than the other. To enable direct comparison, both affordability indexes are standardized using z-scores. Standardization removes scale differences and allows the two series to be interpreted relative to their own historical means and variability.

The resulting joint dynamics plot displays the standardized series from 2001 through 2025, with major macroeconomic events highlighted, including the 2008 housing crash, the COVID shock, and the mortgage-rate spike beginning in mid-2022.

Visual inspection of Appendix I: Figure 26 reveals clear differences in volatility and adjustment speed across the two markets. Home affordability follows a comparatively smooth long-run trajectory, while rent affordability exhibits sharper and more frequent fluctuations. During the 2008 financial crisis, home affordability declines more noticeably, which aligns with the idea that the owner-occupied housing market reacted more directly to the collapse in credit conditions and asset prices. Rent affordability also shifts during this period, but the magnitude of the change is more muted.

A similar pattern emerges during the COVID period. Both affordability measures experience abrupt movements, yet rent affordability displays faster and steeper swings. The pronounced decline in rent affordability around 2021–2022 is particularly striking, as the index falls rapidly below its historical mean before stabilizing at a lower level. In contrast, home affordability recovers more gradually following the initial shock. Overall, the figure suggests that while the two markets share broad macroeconomic movements, their short-run responses differ substantially in both speed and intensity.

5.5.11 Stationarity Assessment and Motivation for Differencing

Before formally analyzing timing relationships, the time-series properties of the affordability measures are evaluated. Augmented Dickey–Fuller (ADF) tests are used to assess whether the level series are stationary. For both home and rent affordability, the tests fail to reject the null hypothesis of a unit root. The test statistic for home affordability is approximately –0.73, well above the 5% critical value, while rent affordability's statistic of about –2.84 lies close to—but does not cross—the rejection threshold.

These results indicate that both affordability measures behave as trending processes rather than stationary ones. This finding is important because correlation and regression analyses applied to non-stationary series may yield misleading results driven by shared trends rather than meaningful dynamic relationships.

To address this issue, both series are differenced. After differencing, the ADF test statistics change substantially. The test statistics for ΔHomeAff and ΔRentAff are approximately –12.9 and –16.4, respectively, far below the 1% critical values. This provides strong evidence that the differenced series are stationary. Because subsequent timing analyses require stationary inputs, the differenced series are used in the remaining steps.

5.5.12 Cross-Correlation Analysis and Lead–Lag Structure

To explore whether one affordability measure systematically leads the other, the cross-correlation function (CCF) is computed using the standardized level series. The CCF measures the correlation between the two series at different lags, allowing for an assessment of temporal ordering. By convention, negative lags correspond to the case in which rent affordability leads home affordability, while positive lags imply the opposite.

As shown in Appendix I: Figure 28, the cross-correlation profile forms a pronounced V-shape, with correlations becoming increasingly negative as lags approach zero. Many correlations at negative lags exceed the significance bands, while correlations at positive lags are weaker and less consistently significant. The strongest negative correlations appear between roughly –10 and 0 months.

This pattern indicates that changes in rent affordability tend to occur earlier than changes in home affordability. In contrast, there is limited evidence that movements in home affordability systematically precede changes in rent affordability. The asymmetry in the correlation structure suggests that the rental market may adjust more rapidly to economic conditions, while the owner-occupied market responds with a delay.

5.5.13 Cross-Lag Regression Evidence

To complement the correlation-based evidence, cross-lag regressions are estimated. These regressions test whether past changes in one affordability measure help explain current changes in the other, while controlling for their own dynamics. This approach provides a more structured way to assess directional relationships beyond simple correlation.

The first specification examines whether lagged changes in rent affordability predict changes in home affordability. The one-month lag carries a large and highly significant negative coefficient (approximately –0.46), indicating that short-run declines in rent affordability are associated with subsequent declines in home affordability. The nine-month lag is also statistically significant and positive, suggesting a secondary, slower adjustment channel. Other lags are not significant, and the model explains roughly 28% of the variation in changes in home affordability.

Reversing the regression direction yields a similar lag structure, with the one-month and nine-month lags remaining significant. However, the overall explanatory power does not increase materially. Taken together, these results reinforce the CCF findings: movements in rent affordability tend to precede movements in home affordability, rather than the reverse.

5.5.14 Interpretation of Lead–Lag Patterns

Taken together, the evidence from the joint dynamics plot, stationarity tests, cross-correlation analysis, and cross-lag regressions points to a consistent conclusion. The two affordability measures share long-run trends but differ markedly in short-run behavior. Rent affordability reacts more quickly to economic shocks and often moves ahead of home affordability by several months.

One plausible explanation lies in the differing adjustment mechanisms across the two markets. Rental prices adjust more frequently and involve lower transaction costs, allowing rent affordability to respond rapidly to changing conditions. In contrast, home affordability is shaped by credit constraints, expectations, and longer planning horizons, which naturally slow

adjustment. As a result, stress in the rental market may serve as an early signal of broader housing affordability pressures.

5.5.15 Descriptive Regression Analysis of Mortgage Rates, Wages, and Affordability

This section explores how broader macroeconomic conditions relate to housing affordability, focusing on the roles of mortgage rates and real wages. The analysis adopts a descriptive regression framework to summarize long-run associations rather than short-run causal effects. Visual evidence from scatterplots is first used to motivate the empirical relationships, followed by regression models and diagnostic checks that assess how these macro-variables align with affordability outcomes in the homeownership and rental markets. Together, the results highlight meaningful differences in how income and financing conditions interact with renting versus owning.

5.5.16 Exploratory Motivation and Scatterplot Evidence

This section examines how mortgage rates and real wages relate to home and rent affordability. The analysis is explicitly descriptive rather than causal. Both affordability measures follow long-run trends, and earlier stationarity tests indicate that mortgage rates and wages are non-stationary in levels. As a result, regression coefficients are interpreted as long-run associations rather than structural or short-run causal effects.

The scatterplots in Appendix I: Figure 29 reveal clear differences across the two markets. Home affordability increases with real wages also shows a positive association with mortgage rates, although the relationship with mortgage rates is relatively dispersed. The correlation matrix confirms these visual impressions: home affordability correlates at approximately 0.50 with mortgage rates and around 0.29 with wages.

Rent affordability displays a markedly different pattern. Its relationship with wages slopes downward, with a strong negative correlation of roughly –0.78. The association with mortgage rates is positive but more modest, around 0.30. These contrasting patterns suggest that income growth and financing conditions interact with renter and owner affordability in fundamentally different ways.

5.5.17 Regression Results for Home Affordability

The first regression models home affordability as a function of mortgage rates and wages. Both coefficients are positive and highly significant. The estimated coefficient on mortgage rates is approximately 0.00335, while the coefficient on wages is about 3.6×10^{-6}. The model explains roughly 43% of the variation in home affordability.

Although a positive association between mortgage rates and home affordability may appear counterintuitive from a short-run perspective, it reflects long-run co-movement between the two series rather than a direct affordability mechanism. The wage coefficient, while small in

magnitude, is precise and consistent with the idea that sustained income growth supports higher home affordability over longer horizons.

5.5.18 Diagnostic Assessment and Model Limitations

Diagnostic checks reveal important limitations of the home-affordability regression. Residual-versus-fitted plots show curvature, indicating that a linear specification does not fully capture the data structure. Residuals over time display slow-moving cycles and a visible distortion during the COVID period. Autocorrelation is particularly strong, with a Durbin–Watson statistic near 0.10, confirming substantial serial dependence.

These patterns suggest that the regression primarily captures shared long-run trends rather than short-run dynamics, reinforcing the descriptive interpretation of the results.

5.5.19 Regression Results for Rent Affordability

Applying the same specification to rent affordability yields notably different results. The coefficient on mortgage rates is positive (approximately 2.3×10^{-4}), while the coefficient on wages is negative (around -2.2×10^{-6}). Both estimates are statistically significant, and the model explains about 62% of the variation in rent affordability.

The sign pattern is economically informative. The negative association between wages and rent affordability suggests that higher wages coincide with lower rent affordability rather than higher. One plausible explanation is that wage growth tends to be strongest in high-demand urban labor markets, which are also the locations where rents have risen most rapidly. In this sense, rising wages and declining rent affordability may reflect the same underlying urban growth pressures.

5.5.20 Residual Diagnostics and Interpretation

Diagnostics for the rent model resemble those of the home-affordability regression, though they are somewhat less severe. Residual plots indicate non-linearity and mild heteroskedasticity, and residuals exhibit persistent cycles with a sharp disruption during the COVID period. Autocorrelation remains strong, with a Durbin–Watson statistic around 0.56, and normality tests reject Gaussian errors.

Overall, the regression results provide a coherent descriptive summary of how affordability aligns with broader economic conditions. Wages show the most consistent associations—positive for home affordability and negative for rent affordability—while mortgage rates are related to both measures through long-run co-movement. The diagnostics underscore that these models should not be interpreted as short-run predictive tools. Nevertheless, the results clarify that income and financing conditions interact with housing affordability in distinct ways across renter and owner segments, reinforcing the need to analyze the two markets separately.

VI. Conclusion & Discussion

The current research carries out an extensive statistical analysis with the aim of highlighting the trend of housing affordability within the United States, spanning the period of the year 2000 to the year 2025, to identify the clear structural divergences that exist between the ownership market and the rental market. The combination of hypothesis testing, inference, time-series analysis, and market assessment seeks to offer a multifaceted perspective on the drivers of the trend and market reaction behaviors to the prevailing economic conditions.

One of the main findings is that the speed and volatility of adjustments in both markets are significantly different: rental affordability adjusts quickly to changes in the economy, showing volatility in the short run; and adjustments in homeowner affordability are slow, with limits imposed by mortgage rates, credit conditions, and price appreciation, which increase volatility in the long run. This observation holds true throughout the analysis. The rental sector is more volatile than the other two sectors, while the homeowner sector, which adjusts slowly, shows large amplitude due to continuous price appreciation. Similarly, the homeowner affordability index, which adjusts slowly, shows large volatility due to mortgage rates, credit conditions, and continuous price appreciation. A recurring distinction in the analysis is that rental affordability exhibits greater short-term sensitivity, while homeownership affordability reflects deeper structural pressures accumulating over time.

The study's time-series analysis reveals another core insight: rental affordability systematically leads homeownership affordability by several months. Cross-correlation patterns show a strong statistically significant negative lag, indicating declines in rental affordability typically precede declines in homeownership affordability. Cross-lagged regression results reinforce this finding, with significant short-term coefficients indicating that rental market pressures reliably predict subsequent fluctuations in the homeownership market. This asymmetry likely stems from differing adjustment frictions—rental prices adjust frequently with low transaction costs, whereas home purchase decisions involve financing constraints, expectations, and slower household planning cycles. Consequently, the rental market serves as an early indicator of broader affordability pressures.

Prediction results further highlight structural divergences across markets. Rental affordability exhibits stable mean reversion, enabling relatively precise and reliable forecasting intervals. In contrast, homeownership affordability displays significant uncertainty, with forecasting intervals rapidly expanding over time. This indicates that homeownership affordability is more sensitive to economic shocks and interest rate changes, making short-term forecasting more challenging. This divergence underscores the rental market's role as a stabilizing force, while the homeownership market remains exposed to volatility risks. It should be noted, however, that bootstrap analysis also supports the statistical validity of the above results. The results from the computation of the confidence interval using the bootstrapping approach indicate a significant negative trend in the long-run affordability of homeownership from the early 2000s to the 2020s.

Moreover, the negative, strongly significant association between mortgage rates and housing affordability holds for thousands of iterations, and it can be suggested that the results are not dependent on distributional assumptions and variability.

Furthermore, descriptive regression analysis offers a macroeconomic perspective on the relationships involved in affordability. The impact of mortgage rates on affordability is significant, although the relationship with real wage rates is complex in nature. These relationships make it apparent that macroeconomic forces act differently on the renting versus the homeowning markets, thus requiring both markets to be separately analyzed instead of being viewed as parallel variables.

Overall, this study demonstrates that while rental and homeownership markets are interconnected, they operate as distinct systems: rental affordability changes more rapidly, exhibits greater volatility, and often acts as a leading indicator; homeownership affordability adjusts more slowly, faces structural constraints, and carries greater uncertainty. These insights hold significant implications for policymakers, economists, and households: monitoring rental affordability provides early warning signals for financial stress, while understanding the structural determinants of homeownership affordability is crucial for designing long-term housing policies.

VII. References

Brookings Institution. (2022). *The widening geography of housing inequality*. Brookings Metro. https://www.brookings.edu

Federal Reserve Board. (2022). *Financial stability report*. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/publications/financial-stability-report.htm

Freddie Mac. (2021). *Housing supply: A growing deficit*. Freddie Mac Economic Research. https://www.freddiemac.com

Joint Center for Housing Studies of Harvard University. (2023). *The state of the nation's housing 2023*. Harvard University. https://www.jchs.harvard.edu

National Association of Realtors. (2023). *Housing affordability index report*. National Association of Realtors. https://www.nar.realtor/research-and-statistics/housing-statistics/housing-affordability-index

Pew Research Center. (2024). *Homeownership and affordability trends among U.S. adults*. Pew Research Center. https://www.pewresearch.org

U.S. Bureau of Economic Analysis. (2025). *Disposable personal income: Per capita (A229RC0)* [Data set]. Federal Reserve Bank of St. Louis (FRED). https://fred.stlouisfed.org/series/A229RC0

U.S. Bureau of Economic Analysis. (2025). *SAINC51 State annual disposable personal income summary: Disposable personal income, population, and per capita DPI*. https://www.bea.gov (Accessed December 3, 2025)

U.S. Bureau of Labor Statistics. (2025). *Consumer Price Index for All Urban Consumers: Rent of Primary Residence (CUSR0000SEHA)* [Data set]. Federal Reserve Bank of St. Louis (FRED). https://fred.stlouisfed.org/series/CUSR0000SEHA

Freddie Mac. (2025). *30-Year Fixed Rate Mortgage Average in the United States (MORTGAGE30US)* [Data set]. Federal Reserve Bank of St. Louis (FRED). https://fred.stlouisfed.org/series/MORTGAGE30US

U.S. Department of Housing and Urban Development. (2023). *Worst case housing needs: 2023 report to Congress*. https://www.huduser.gov

Urban Institute. (2022). *Housing affordability and supply in the United States*. Urban Institute. https://www.urban.org

VIII. Appendix

Appendix I

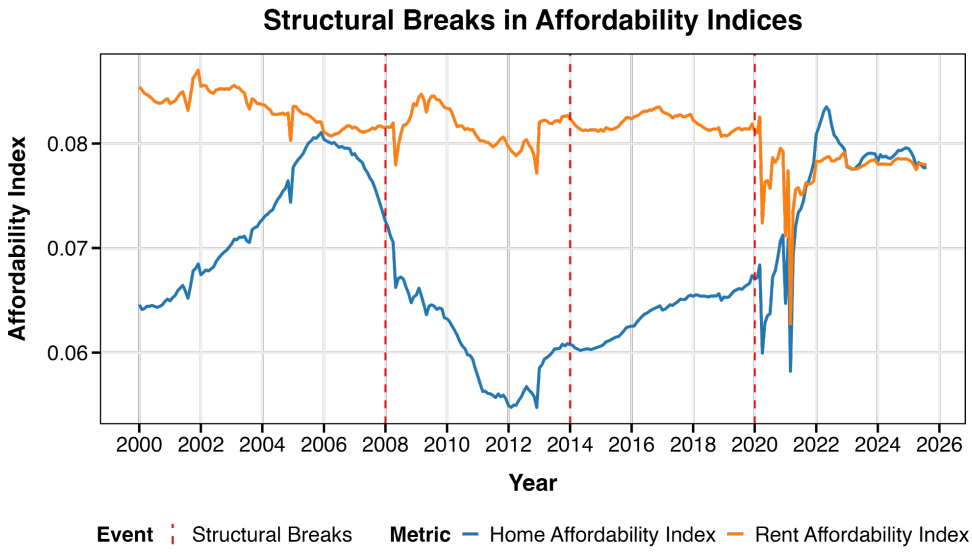


Figure1: Structural Breaks in Housing Affordability and Rent Affordability

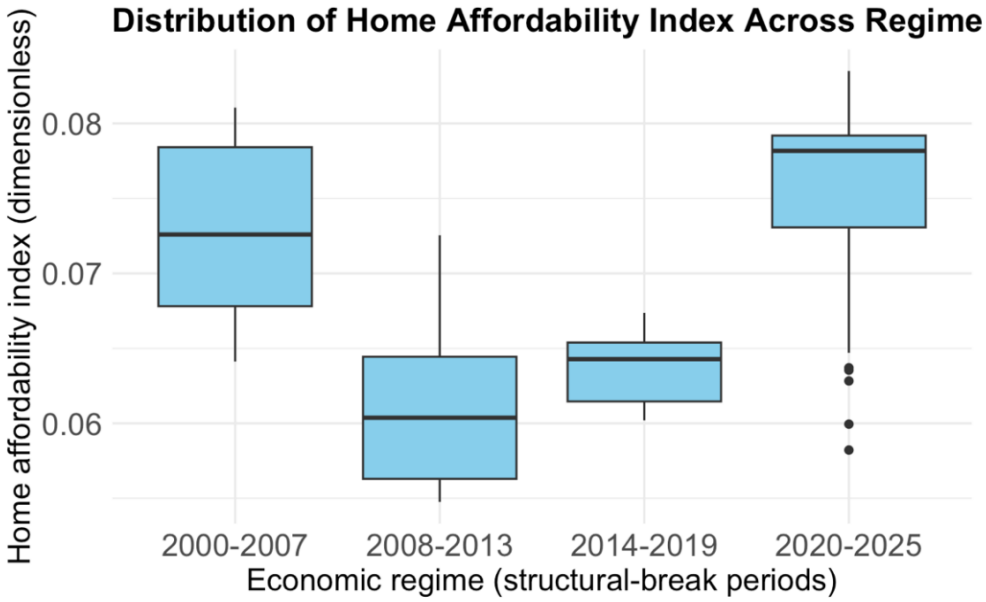


Figure2: Boxplot of Home Affordability Index across Regime

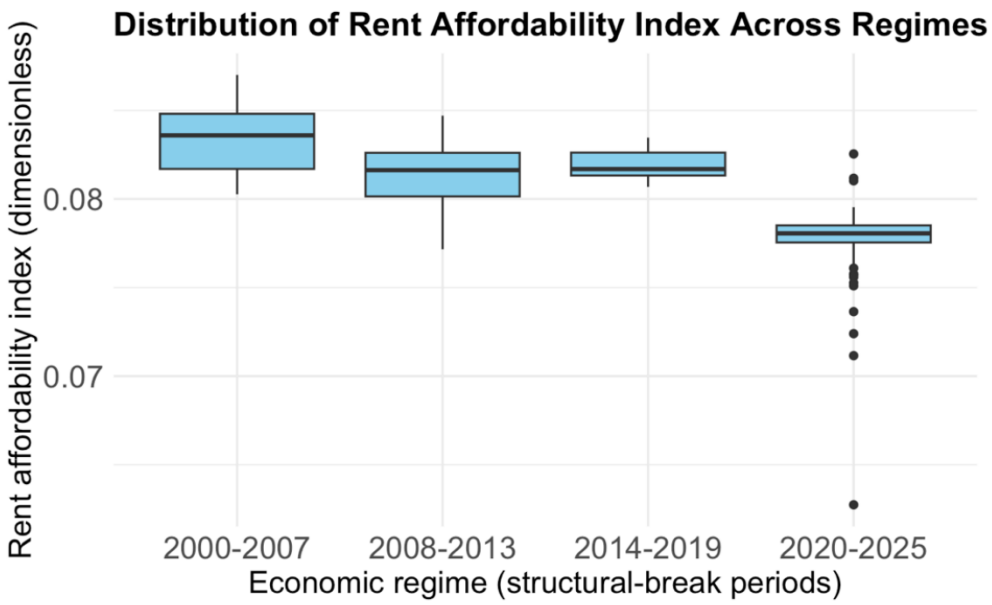


Figure3: Boxplot of Rent Affordability Distribution Across Regime

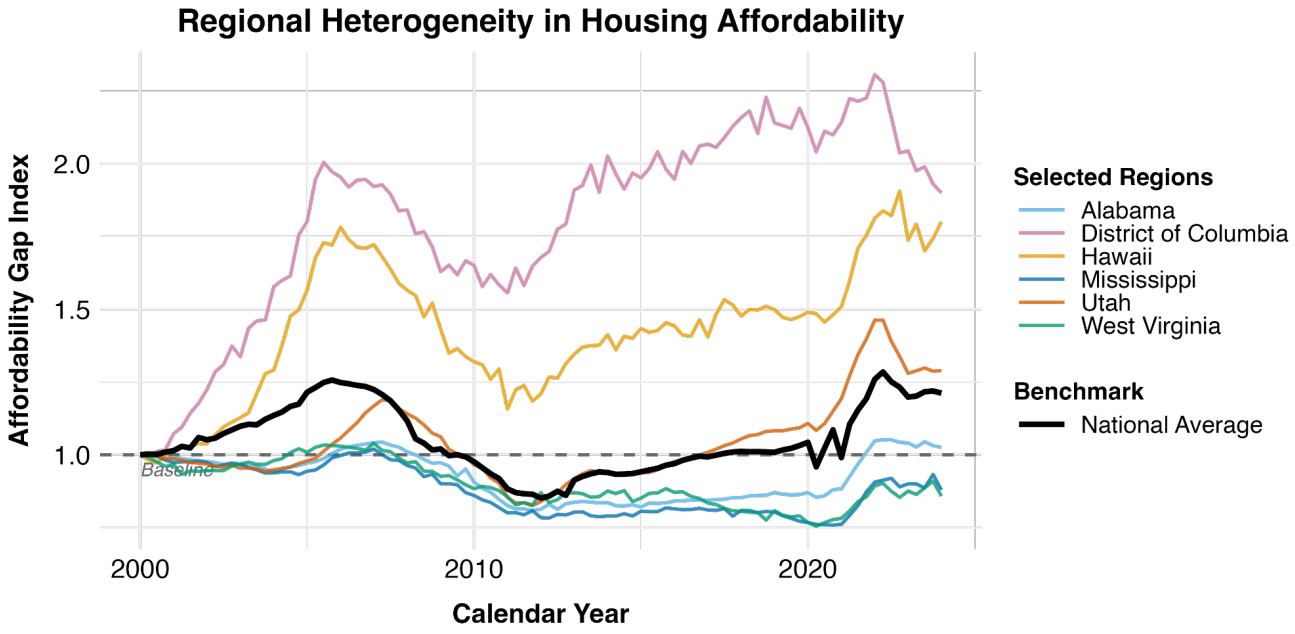


Figure 4: Regional Heterogeneity

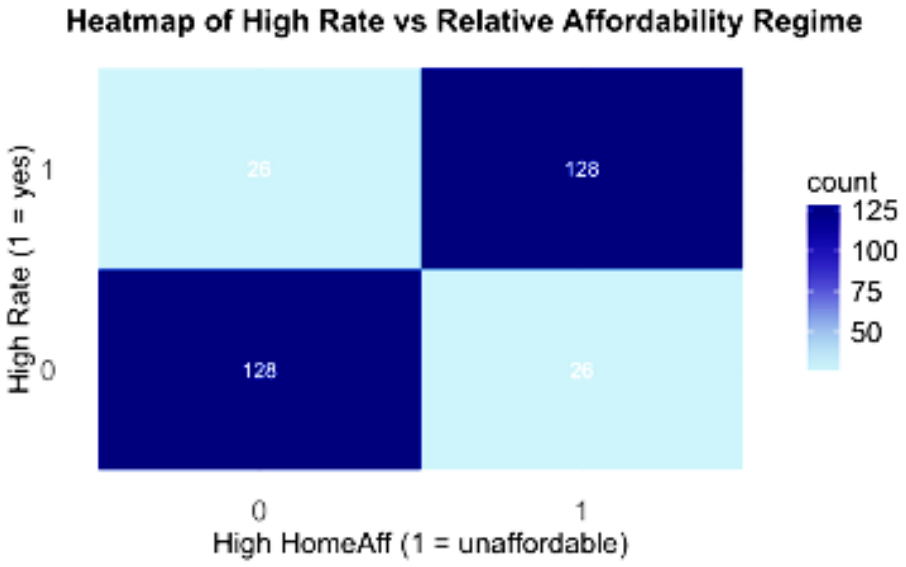


Figure 5: Heatmap of Mortgage Rate vs. Relative Affordability Regime

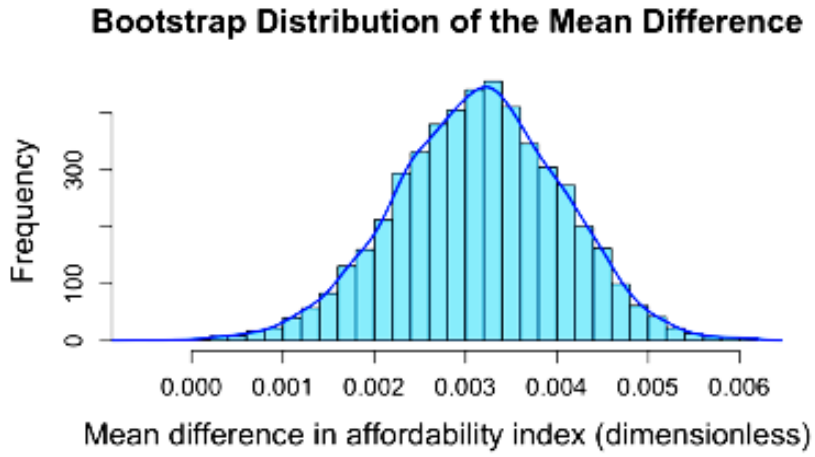


Figure 6: Bootstrap Distribution of the Mean Difference

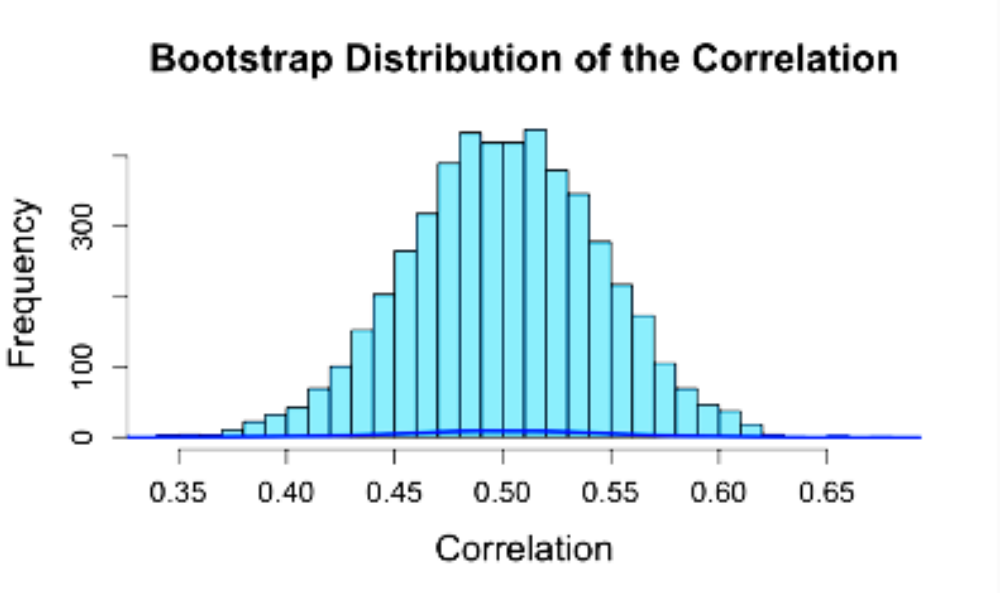


Figure 7: Bootstrap Distribution of the Correlation

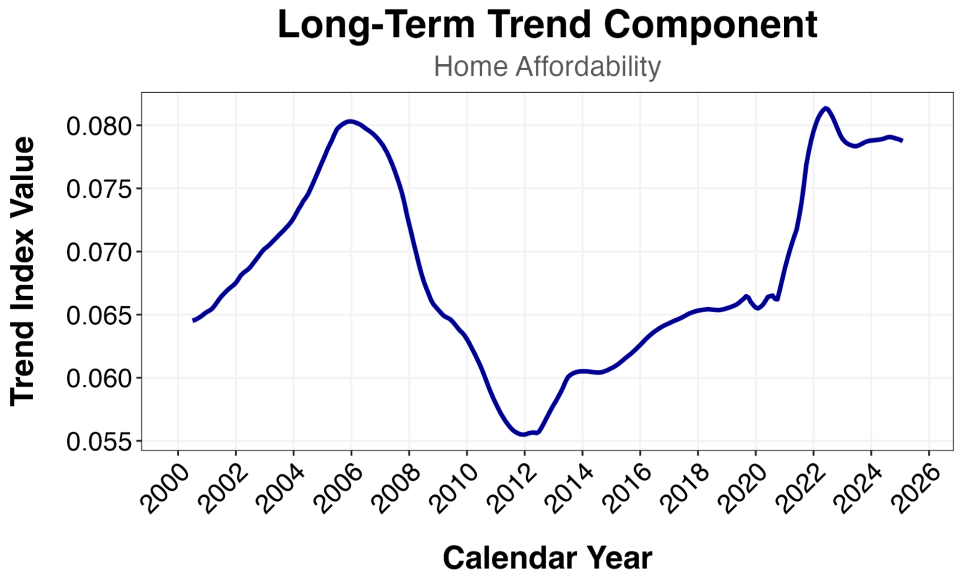


Figure 8: Trend component of Home Affordability

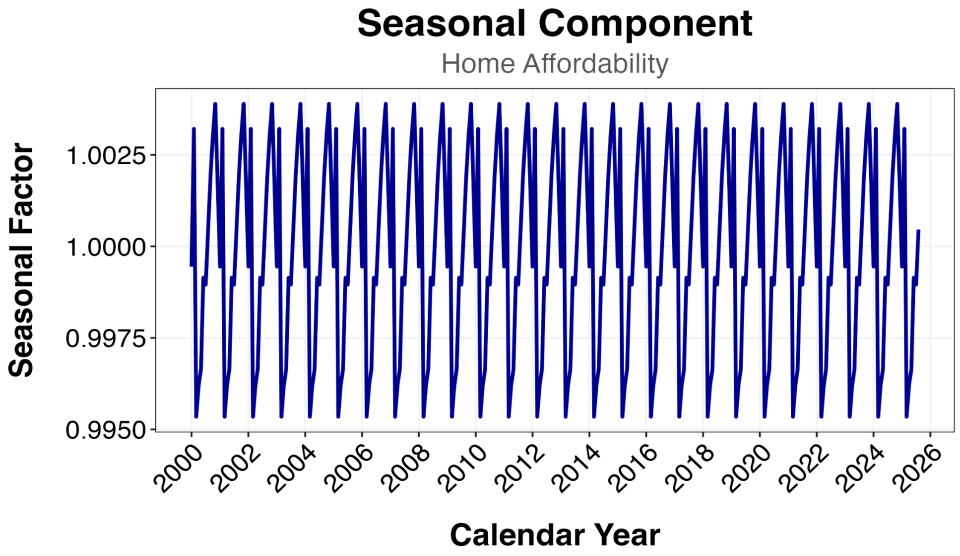


Figure 9: Seasonal Component of Home Affordability

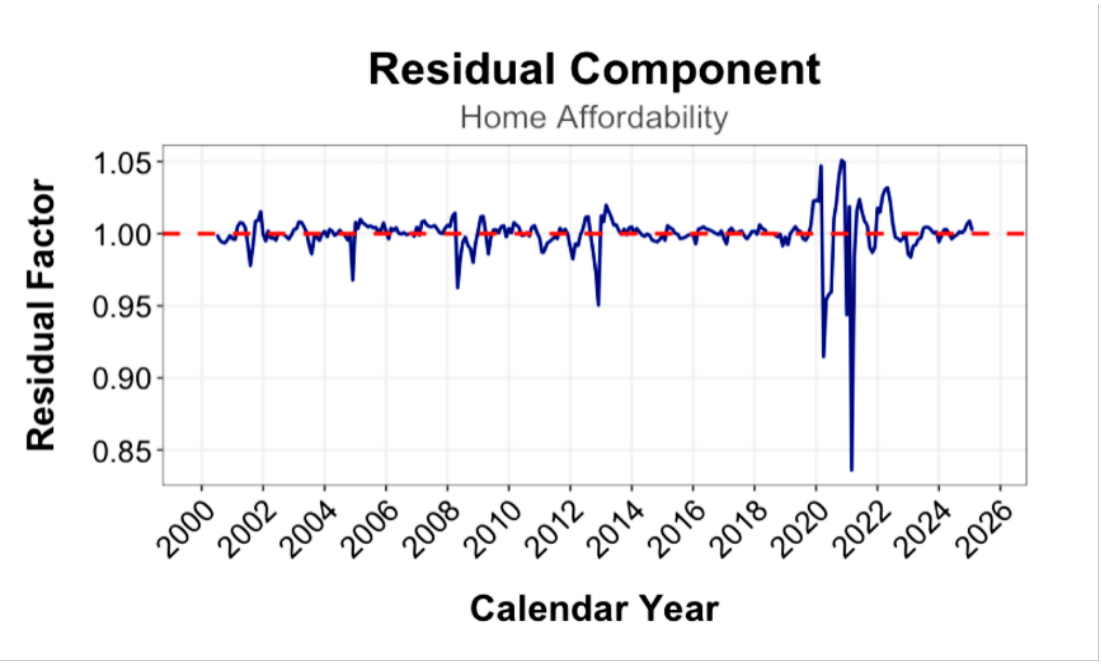


Figure 10: Residual Component of Home Affordability

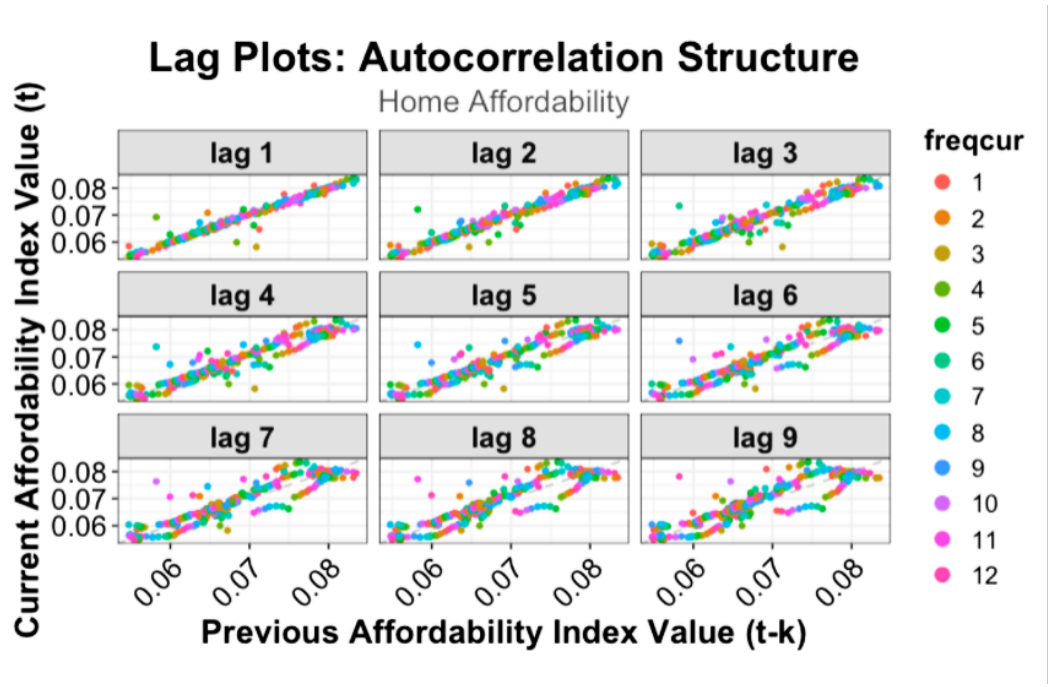


Figure 11: Lag Plots of Home Affordability

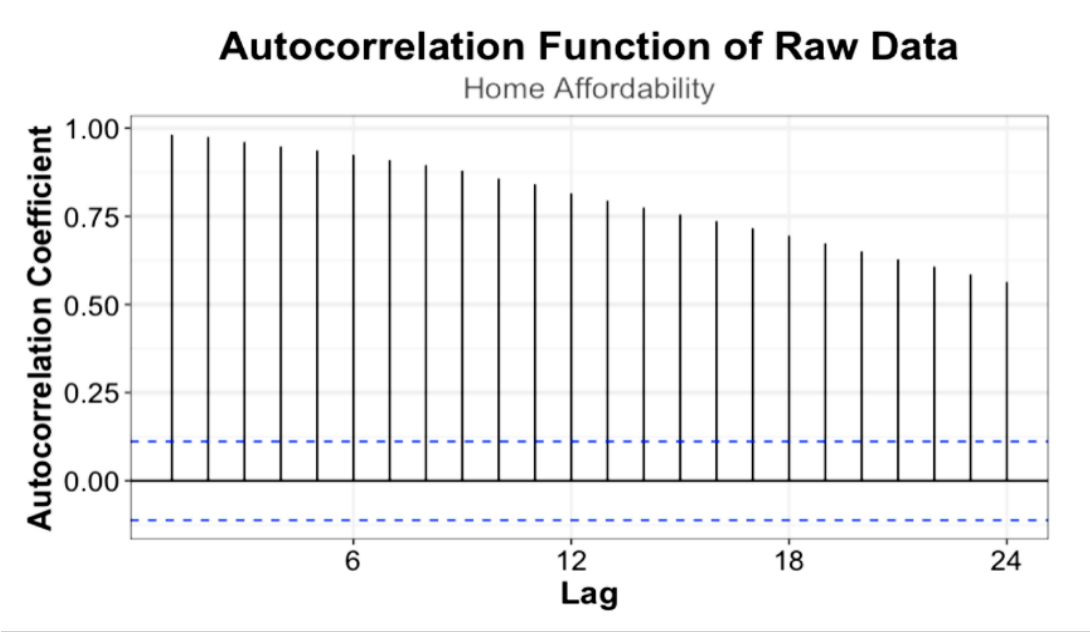


Figure 12: ACF of Raw Data

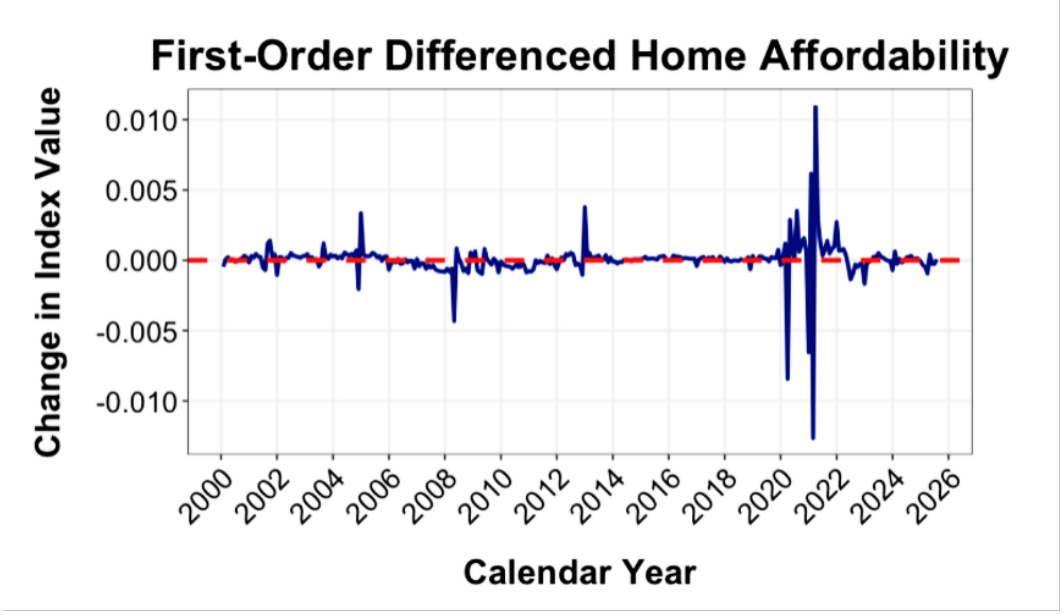


Figure 13: First-Order Differenced Home Affordability

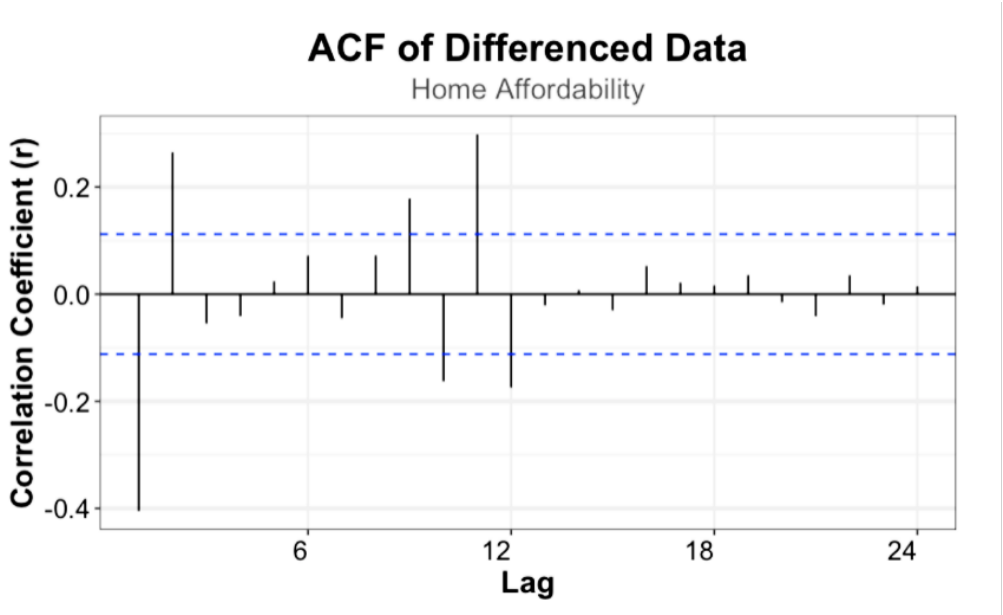


Figure 14: ACF of Differenced Home Affordability Data

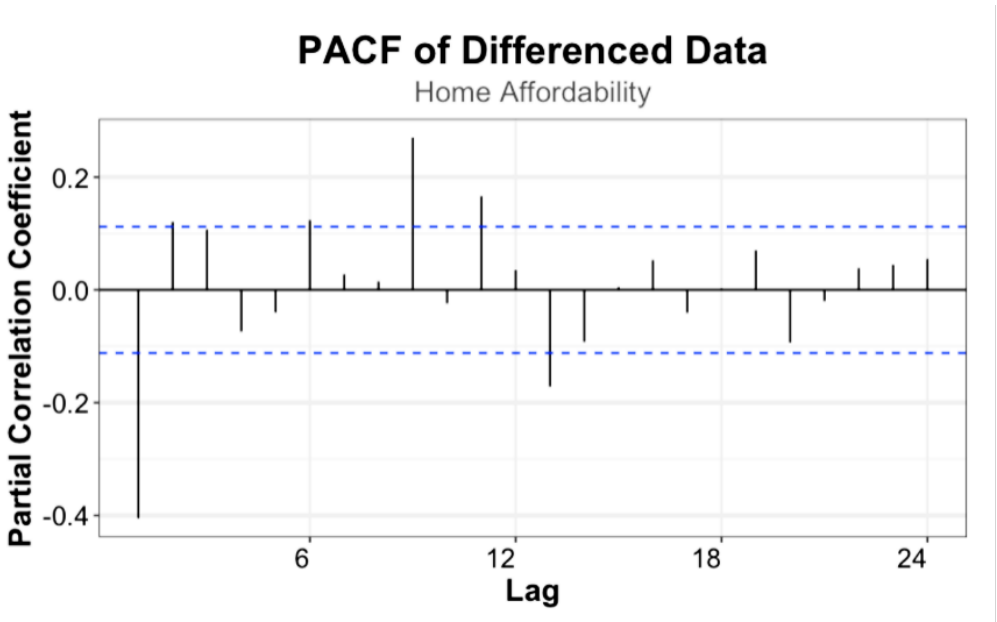


Figure 15: PACF of Differenced Home Affordability Data

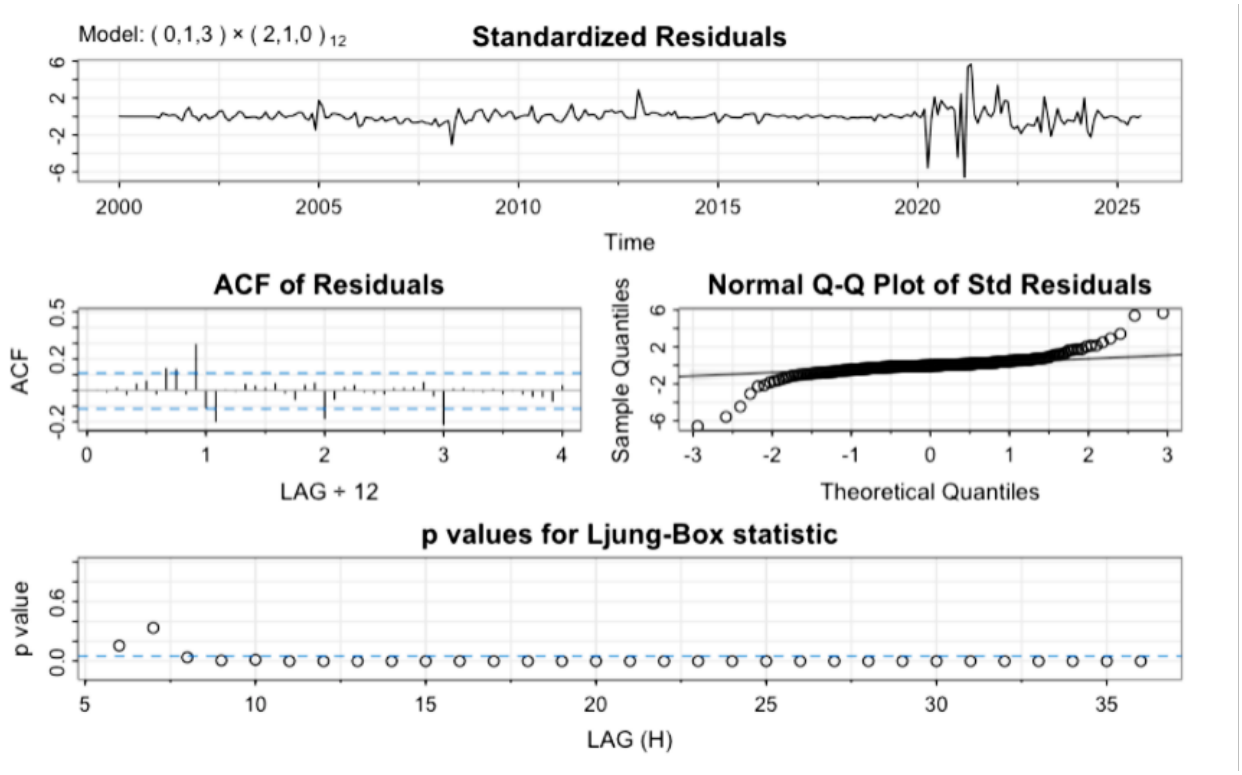


Figure 16: Diagnostics of Automated Model

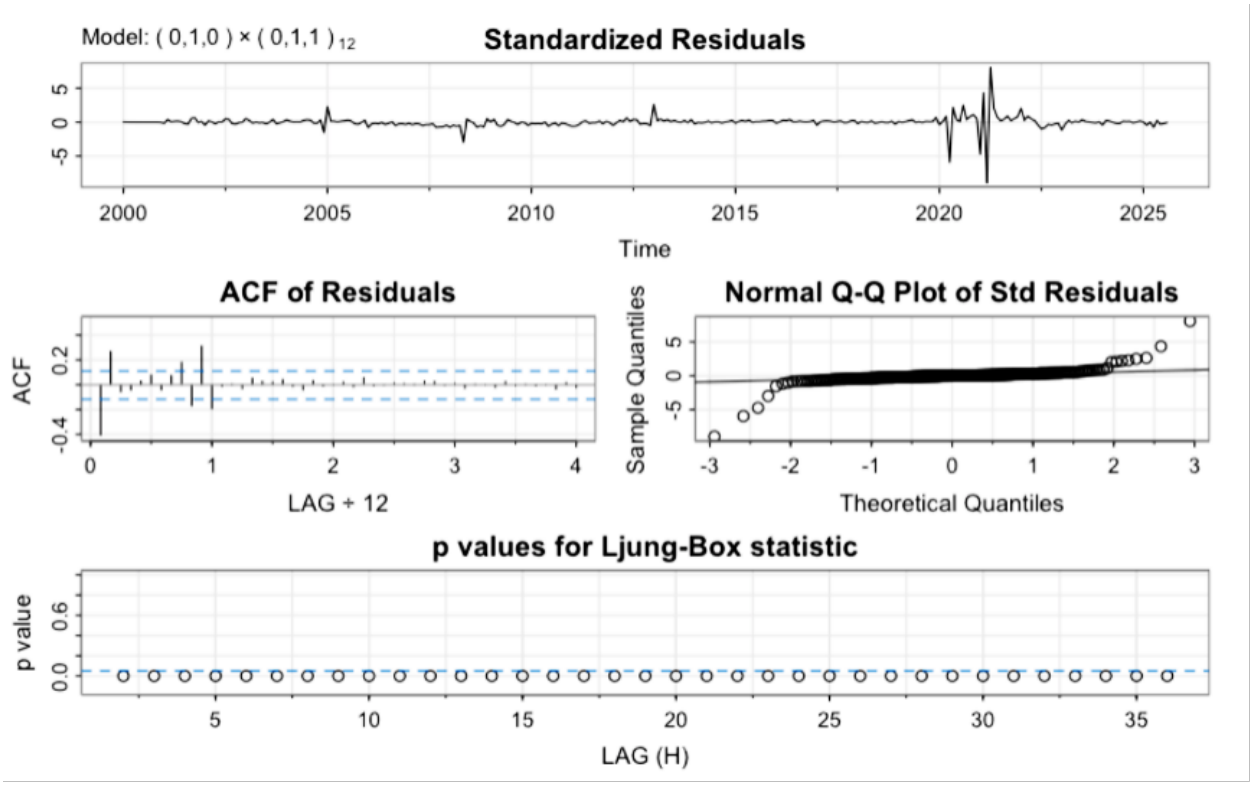


Figure 17: Diagnostics of Manual Model

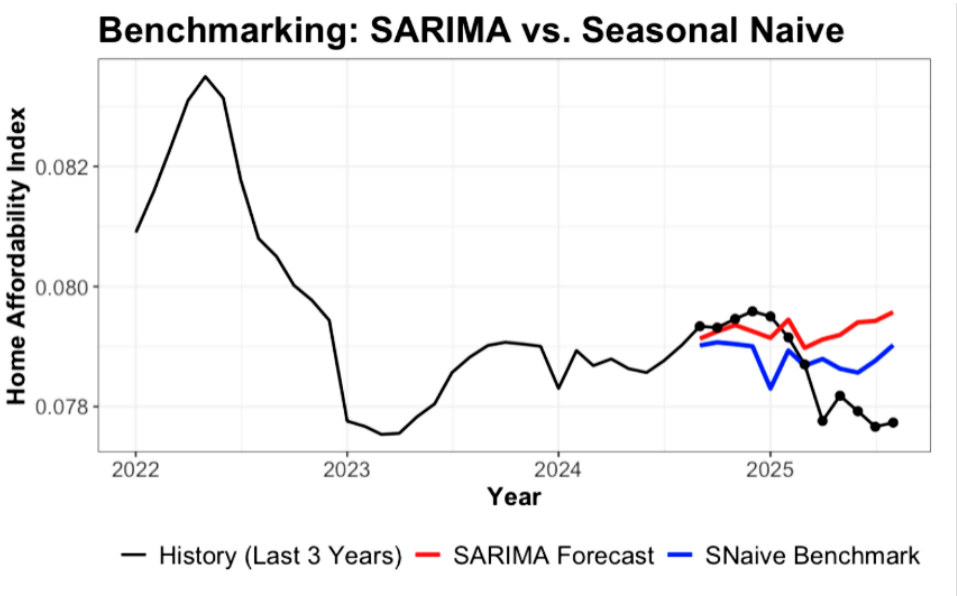


Figure 18: Benchmarking - SARIMA vs. SNaive

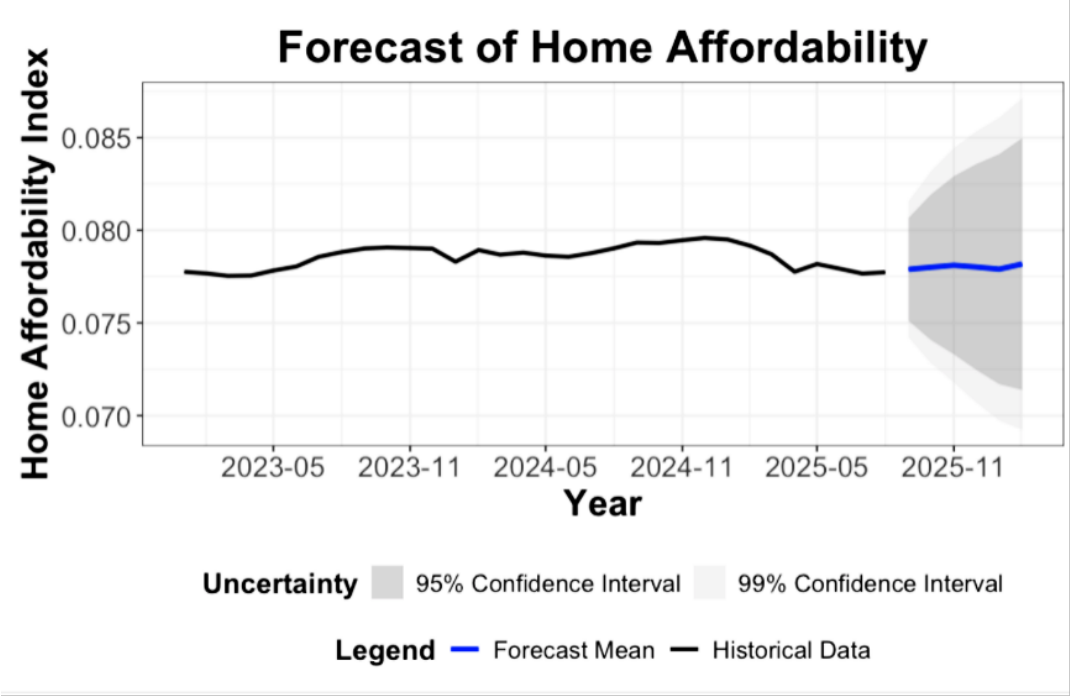


Figure 19: Forecast of Home Affordability

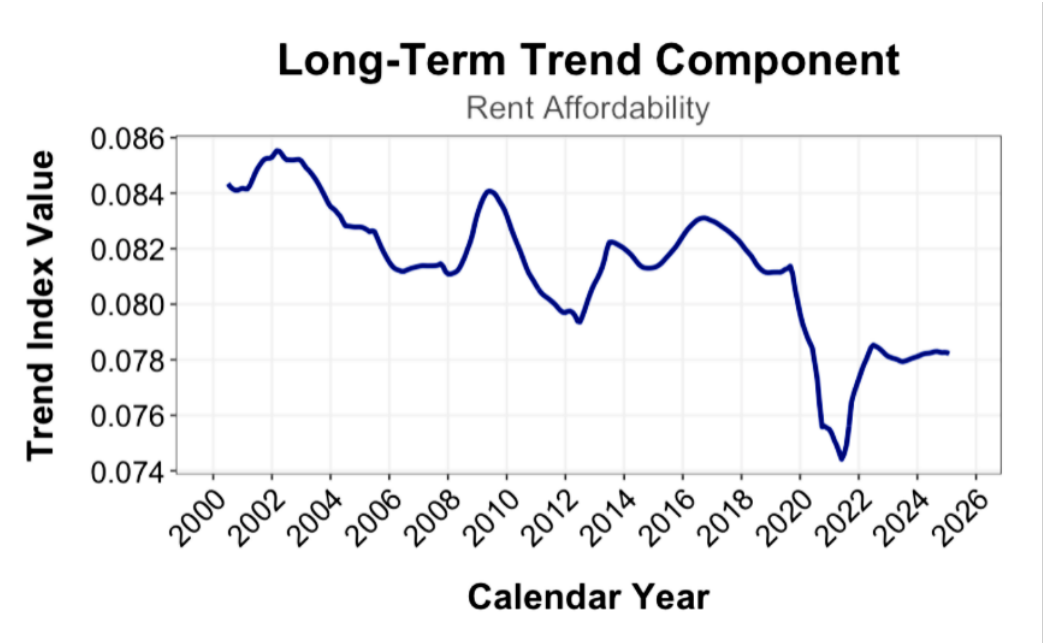


Figure 20: Trend component of Rent Affordability

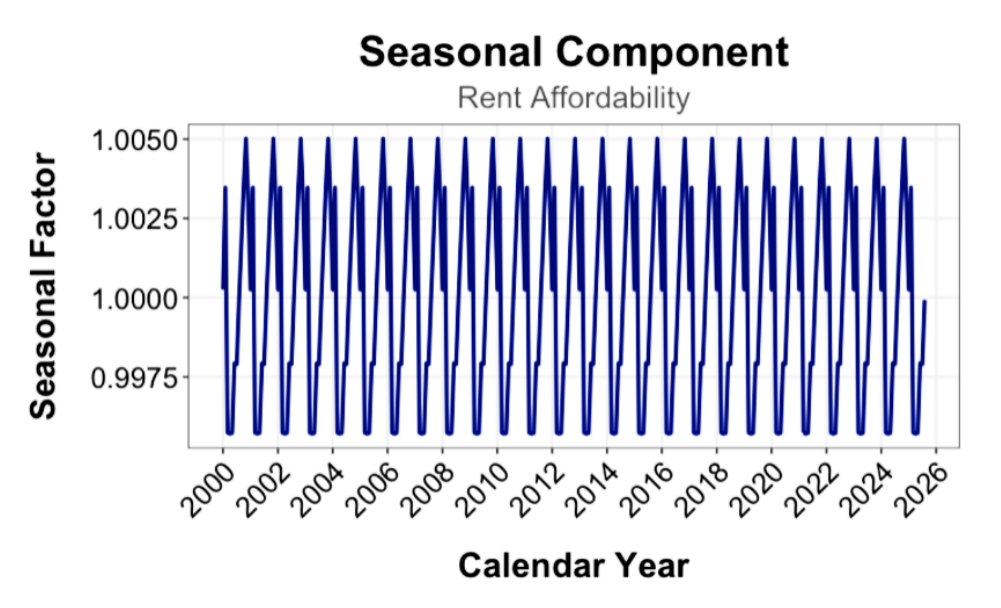


Figure 21: Seasonal component of Rent Affordability

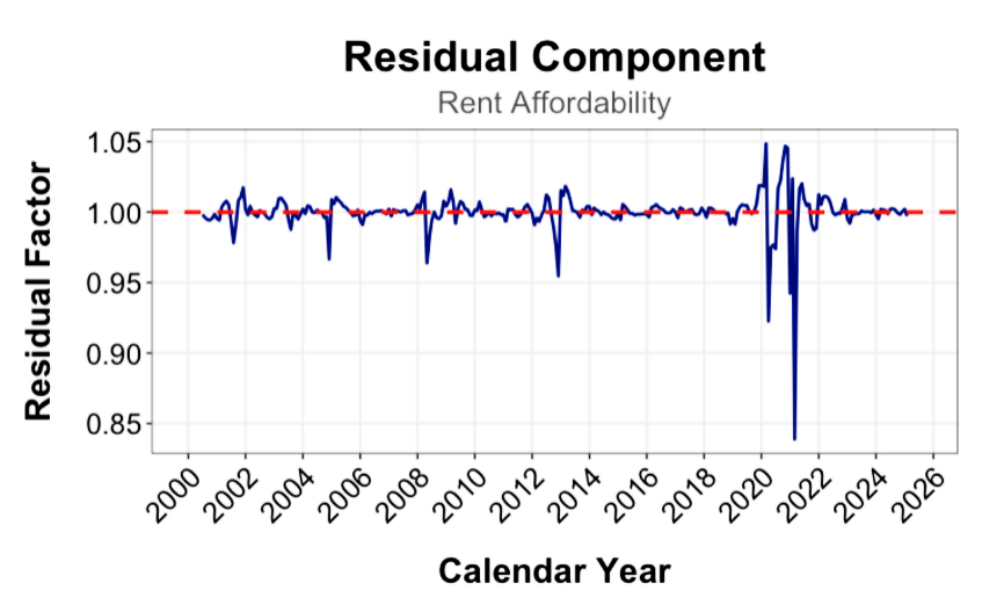


Figure 22: Residual component of Rent Affordability

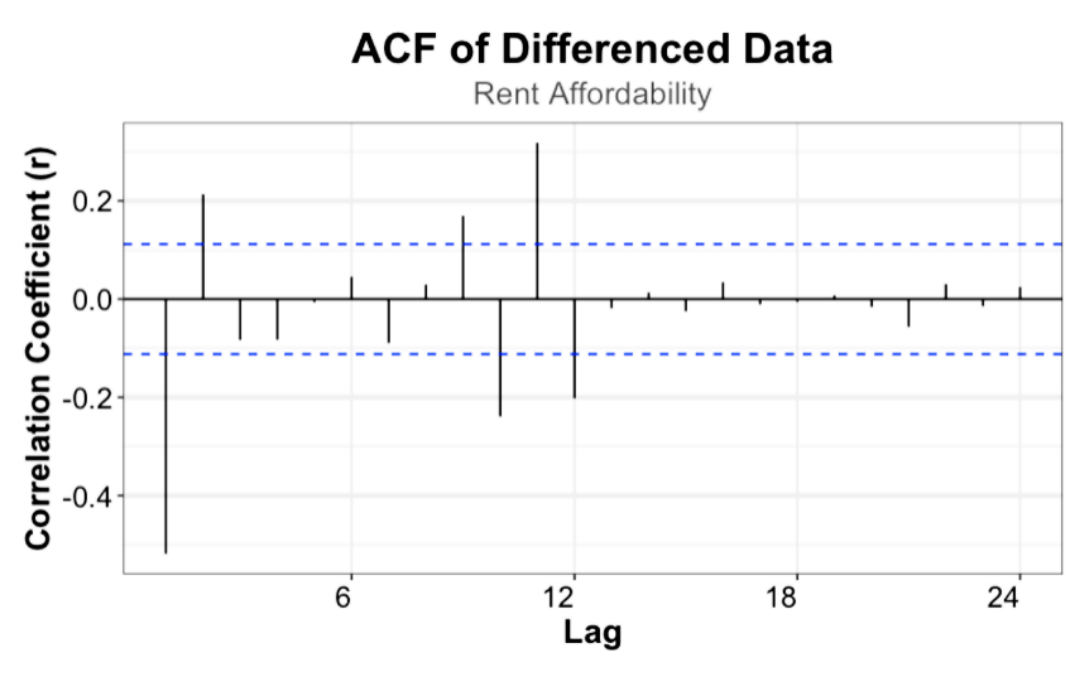


Figure 23: ACF of Differenced Rent Affordability Data

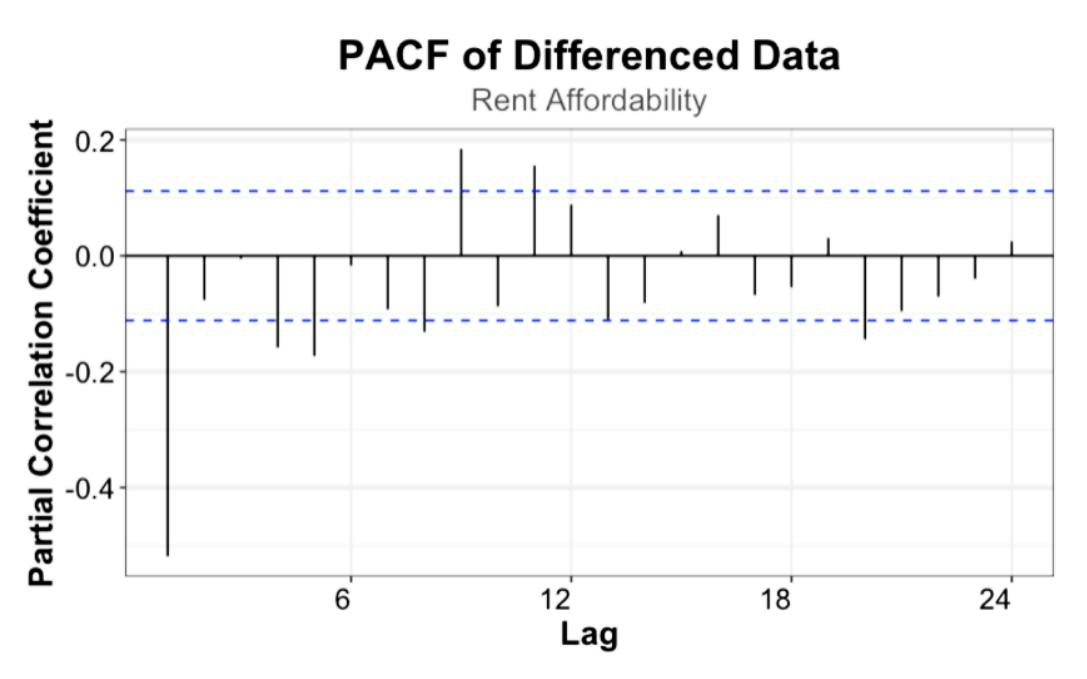


Figure 24: PACF of Differenced Rent Affordability Data

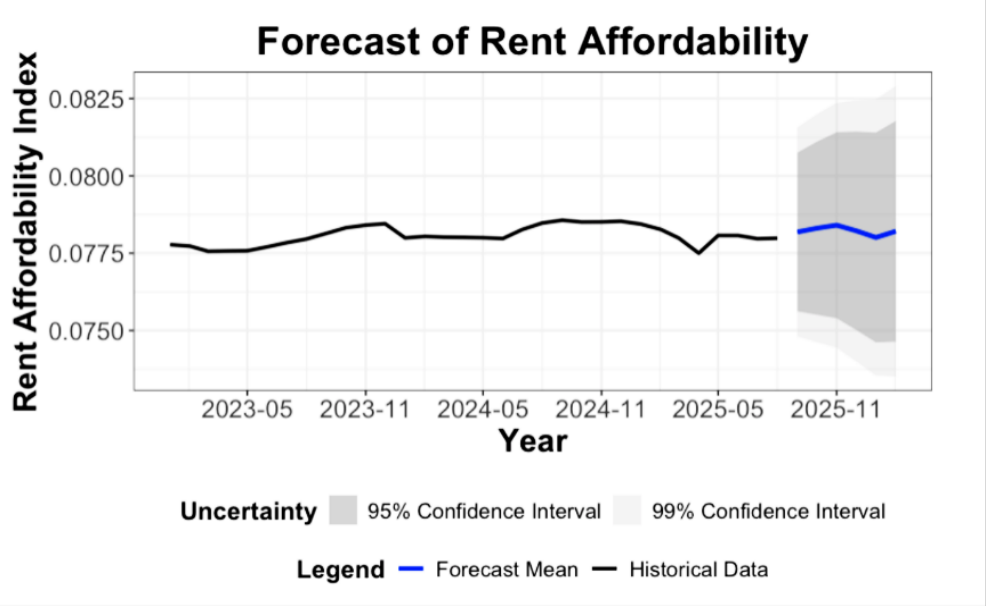


Figure 25: Forecast of Rent Affordability

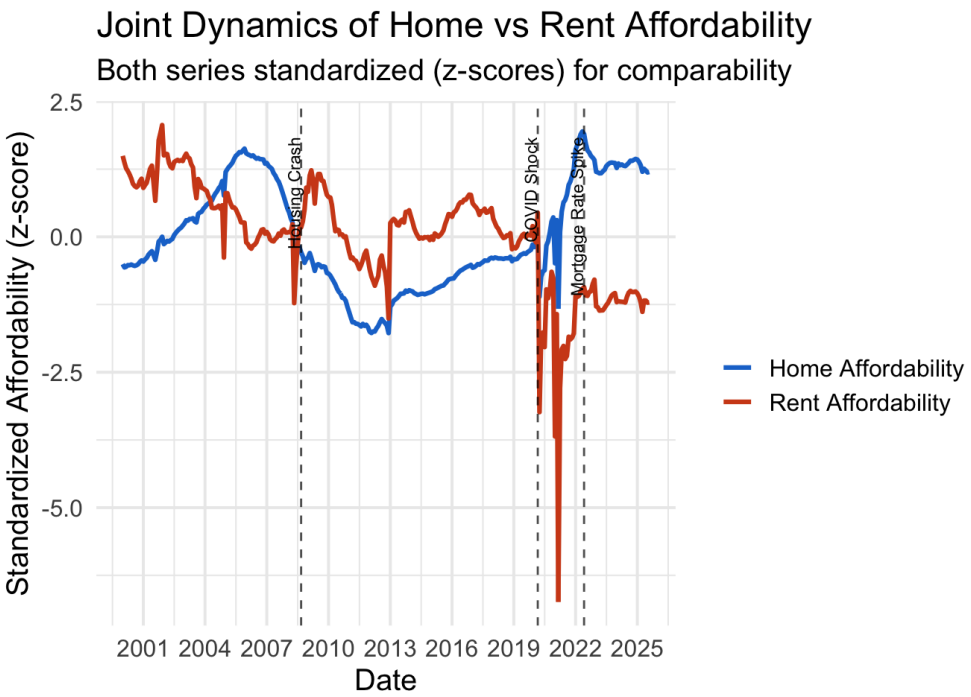


Figure 26: Joint Dynamics of Standardized Home and Rent Affordability with Major Economic Events

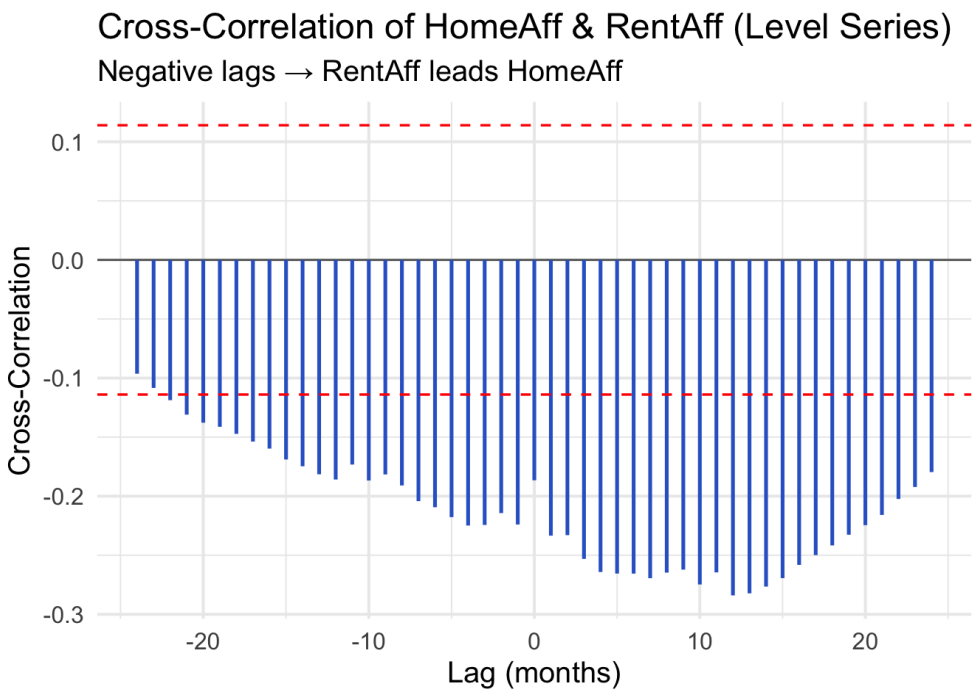


Figure 27: ADF Test Results for Level and Differenced Affordability Series

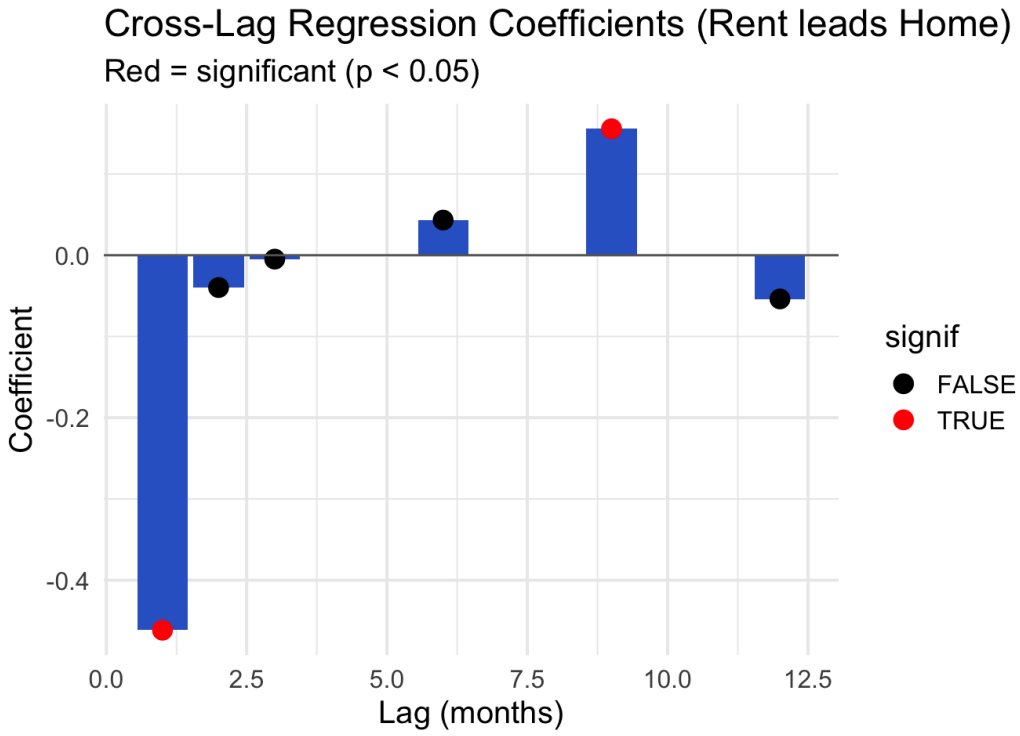
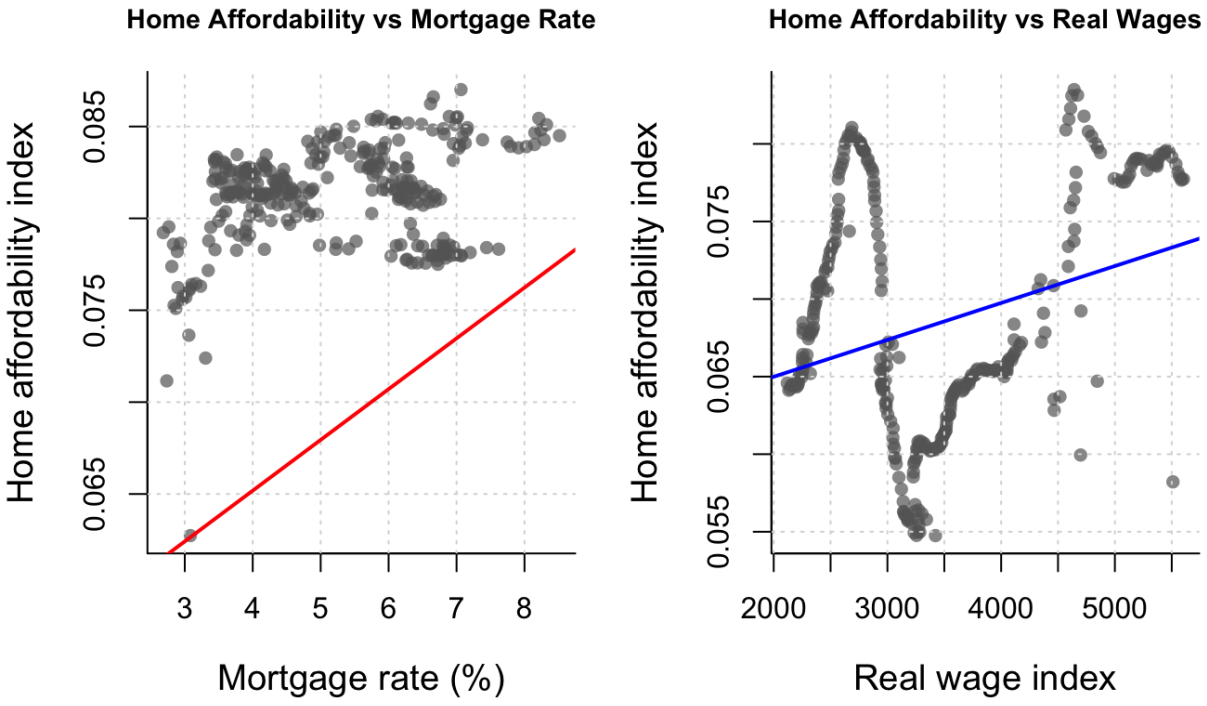


Figure 28: Cross-Correlation Between Home and Rent Affordability (Negative Lags Indicate Rent Leading)

Home Affordability vs Mortgage Rate and Real Wages



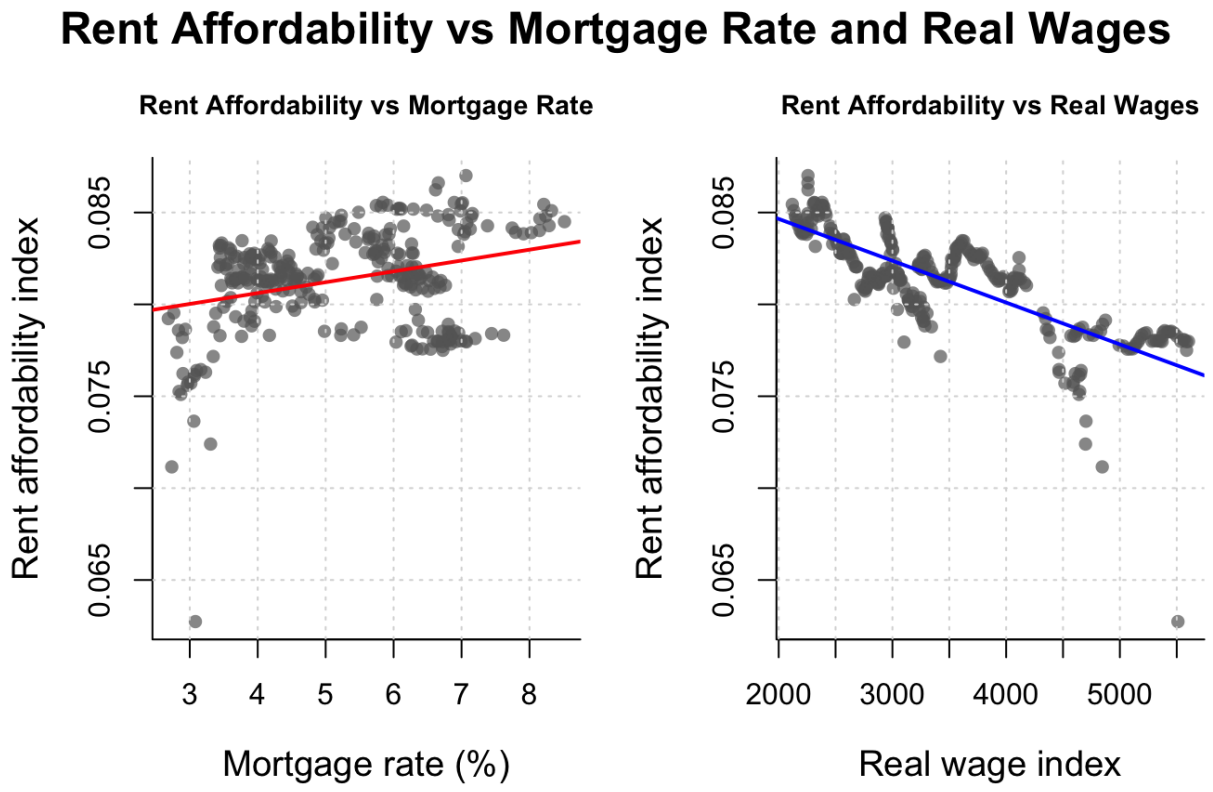


Figure 29: Scatterplots of Home and Rent Affordability Against Mortgage Rates and Real Wages

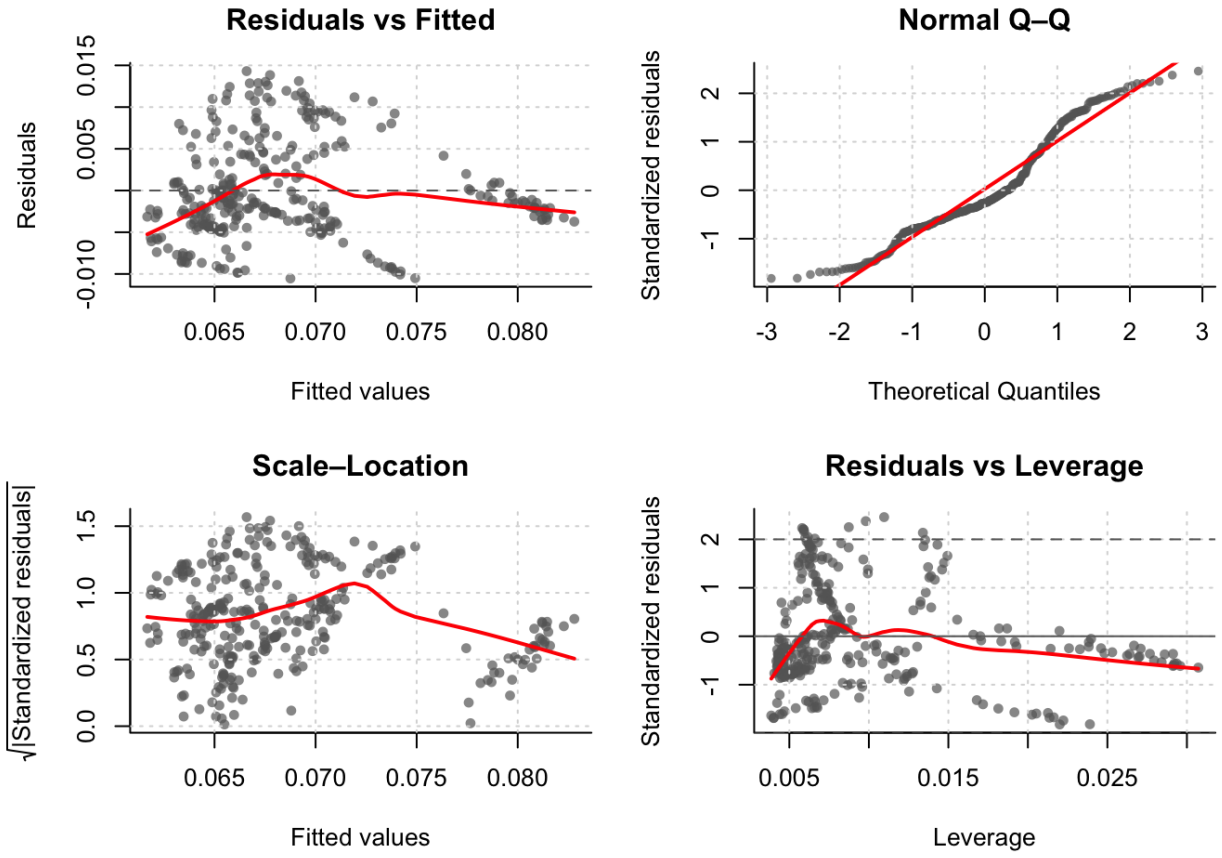


Figure 30: Regression Results for Home Affordability

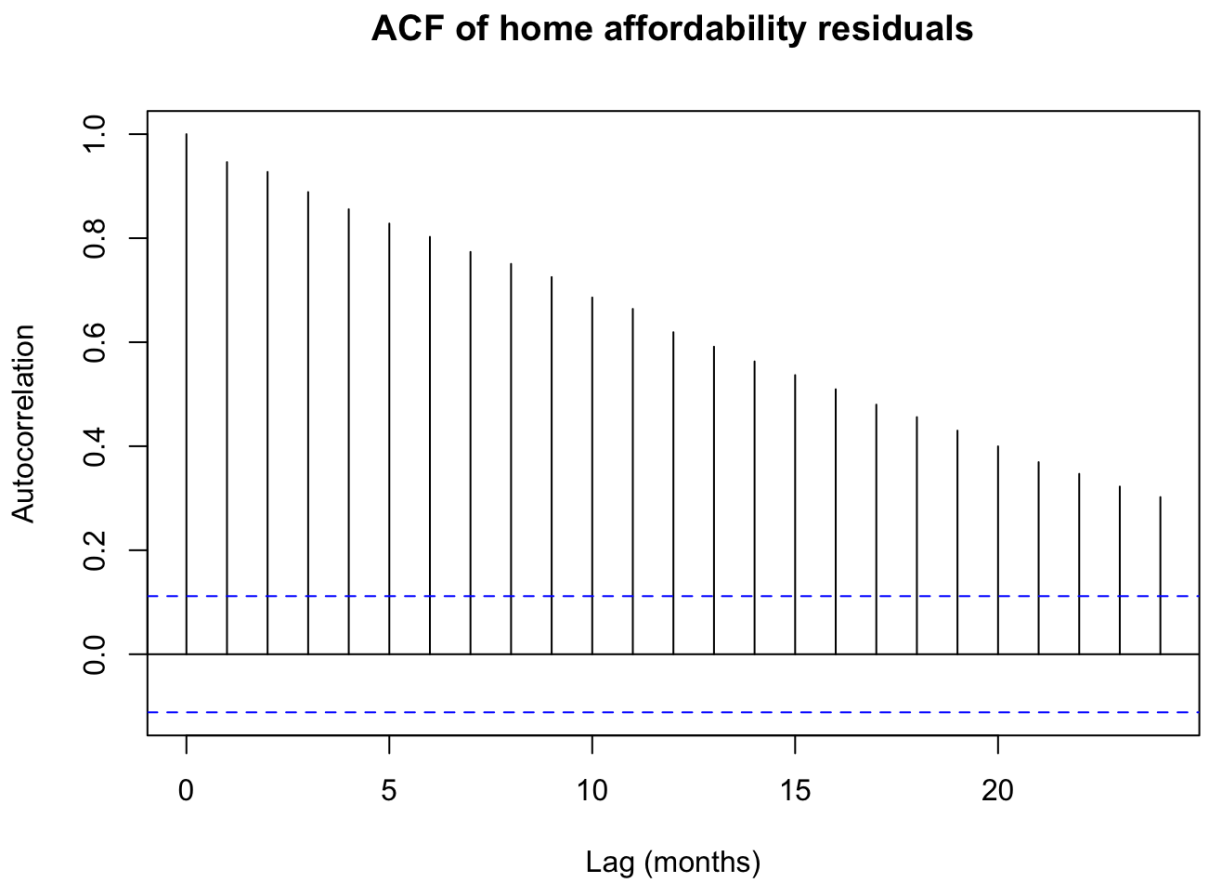


Figure 31: Autocorrelation and Time-Series Diagnostics for Home Affordability Residuals

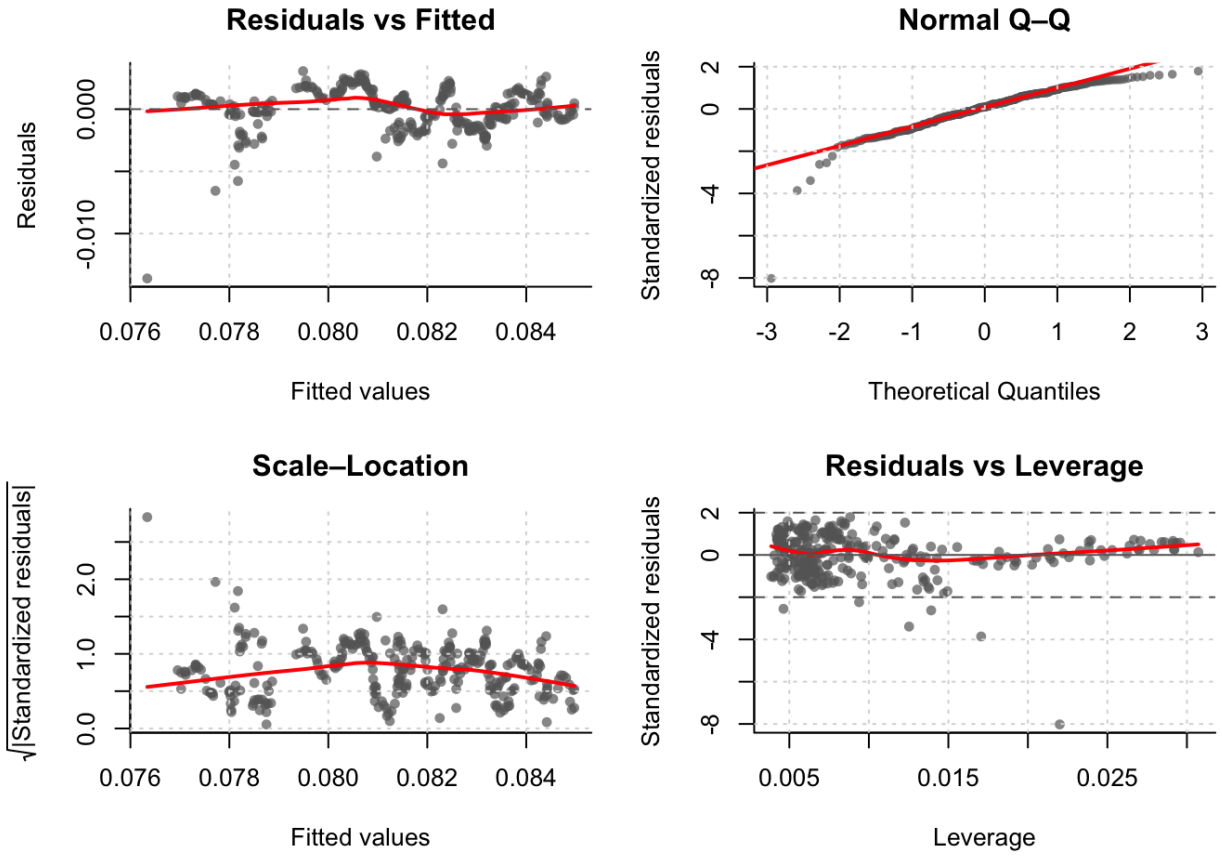


Figure 32: Regression Results for Rent Affordability

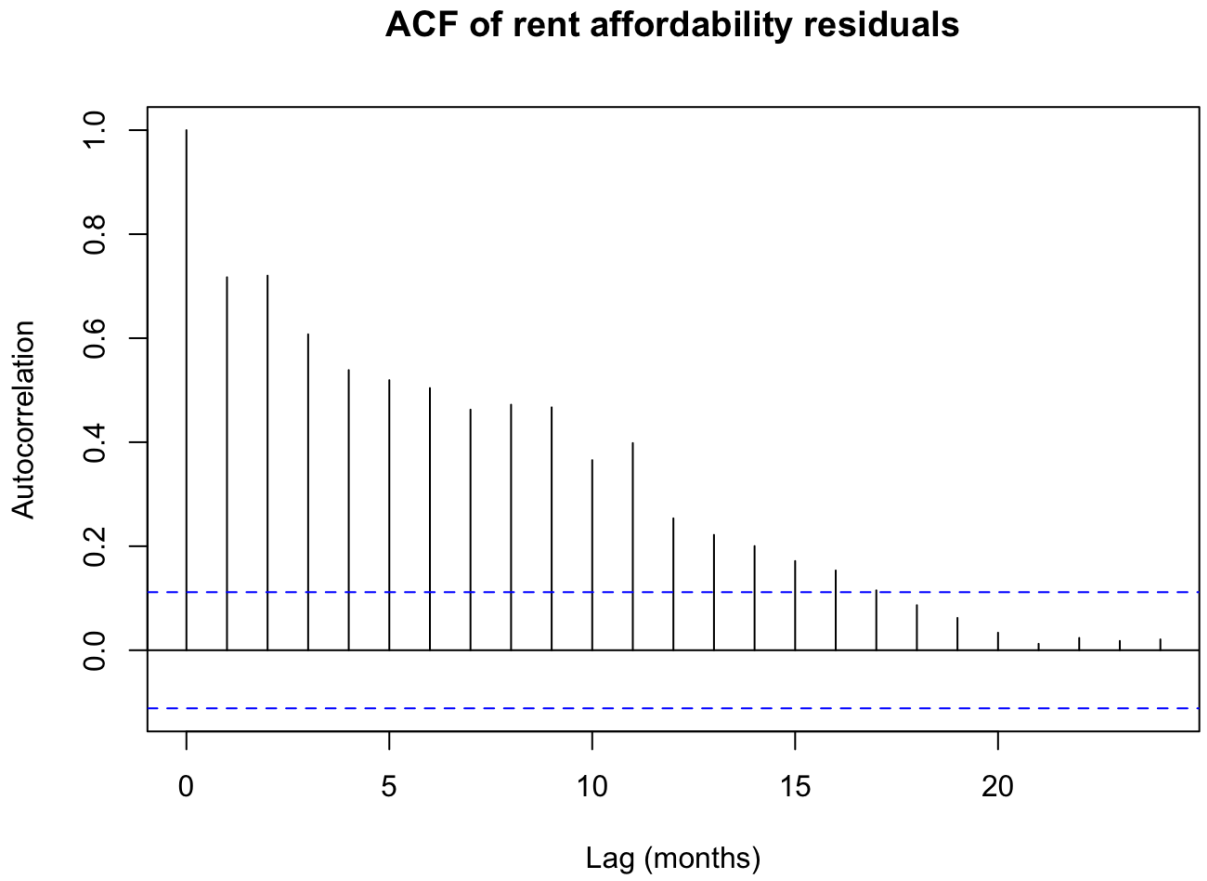


Figure 33: Residual Diagnostics for Rent Affordability Regression

Appendix II

All the data in this report can be accessed through Google Drive:
https://drive.google.com/drive/folders/1fS83EaC66jKIcK-ulMvlH0fIWR7BdYfw?usp=sharing

Appendix III

The complete source code of this report is as follows:

DSAN5100 Final Project - R

Jing Tan Yanzhen Gao Yanmin Gui Jiayuan Gong

Table of contents

Understanding the Housing Affordability Crisis: National Trends, Regional Disparities, and Economic Drivers in the U.S.

(DSAN 5100-01)

Implementation and Reproducibility Statement

All data cleaning, statistical testing, and model estimation were conducted within a reproducible computational environment. The analytical workflow and results can be fully replicated using the original publicly available data. This design ensures that research conclusions are not only statistically valid but also methodologically transparent and reproducible.

1. Data Collection

```
# Load required packages
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.1     v stringr   1.5.2
v ggplot2   4.0.0     v tibble    3.3.0
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.1.0
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)

# 1. Read the original FHFA master CSV
# Make sure "hpi_master.csv" is in your current working directory
hpi_raw <- read_csv("data_part/hpi_master.csv", show_col_types = FALSE)

# 2. Create a proper Date column for monthly and quarterly data

# Helper function: build a Date for monthly data (year-month-01)
make_date_monthly <- function(year, month) {
  as.Date(sprintf("%d-%02d-01", year, month))
}

# Helper function: build a Date for quarterly data (use quarter-end month)
# Here we set Q1 -> March, Q2 -> June, Q3 -> September, Q4 -> December
make_date_quarterly <- function(year, quarter) {
  month <- quarter * 3
  as.Date(sprintf("%d-%02d-01", year, month))
}

hpi <- hpi_raw %>%
  mutate(
    date = case_when(
      frequency == "monthly"   ~ make_date_monthly(yr, period),
      frequency == "quarterly" ~ make_date_quarterly(yr, period),
      TRUE ~ NA_Date_
    )
  )

# only keep observations from year 2000 onward
hpi_2000 <- hpi %>%
  filter(yr >= 2000)

# 3. National monthly SA series (purchase-only)

national_monthly_sa <- hpi_2000 %>%
```

```
  filter(
    level      == "USA or Census Division",
    place_name == "United States",
    frequency  == "monthly",
    hpi_flavor == "purchase-only"
  ) %>%
  arrange(date) %>%
  select(
    place_name, place_id, level, hpi_type, hpi_flavor,
    frequency, yr, period, date, index_sa
  )

# Write to CSV
write_csv(national_monthly_sa, "data_part/fhfa_us_monthly_sa.csv")

# 4. National quarterly SA series

national_quarterly_sa <- hpi_2000 %>%
  filter(
    level      == "USA or Census Division",
    place_name == "United States",
    frequency  == "quarterly",
    hpi_flavor == "purchase-only"
  ) %>%
  arrange(date) %>%
  select(
    place_name, place_id, level, hpi_type, hpi_flavor,
    frequency, yr, period, date, index_sa
  )

write_csv(national_quarterly_sa, "data_part/fhfa_us_quarterly_sa.csv")

# 5. Helper function: quarterly SA series for one state

get_state_quarterly_sa <- function(df, state_code) {
  df %>%
    filter(
      hpi_flavor == "purchase-only",
      frequency  == "quarterly",
      level      == "State",
      place_id   == state_code,
      yr         >= 2000
    ) %>%
    arrange(date) %>%
    select(
      place_name, place_id, level, hpi_type, hpi_flavor,
      frequency, yr, period, date, index_sa
    )
}

# 6. Extract the six states / regions of interest
# Hawaii (HI), West Virginia (WV), Utah (UT), Alabama (AL),
# District of Columbia (DC), Mississippi (MS)
```

```r
state_hi_sa <- get_state_quarterly_sa(hpi, "HI")
state_wv_sa <- get_state_quarterly_sa(hpi, "WV")
state_ut_sa <- get_state_quarterly_sa(hpi, "UT")
state_al_sa <- get_state_quarterly_sa(hpi, "AL")
state_dc_sa <- get_state_quarterly_sa(hpi, "DC")
state_ms_sa <- get_state_quarterly_sa(hpi, "MS")

# 7. Write each state+frequency combination to its own CSV

write_csv(state_hi_sa, "data_part/fhfa_hi_quarterly.csv")
write_csv(state_wv_sa, "data_part/fhfa_wv_quarterly.csv")
write_csv(state_ut_sa, "data_part/fhfa_ut_quarterly.csv")
write_csv(state_al_sa, "data_part/fhfa_al_quarterly.csv")
write_csv(state_dc_sa, "data_part/fhfa_dc_quarterly.csv")
write_csv(state_ms_sa, "data_part/fhfa_ms_quarterly.csv")
```

2. Data Cleaning

2.1 Main Story Data

```r
## 0. Read all raw CSV files

## FHFA national monthly SA house price index
fhfa <- read.csv("data_part/fhfa_us_monthly_sa.csv",
                 stringsAsFactors = FALSE)

## Disposable Personal Income (DPI) from FRED
## Column 'A229RC0' will be divided by 12
dpi  <- read.csv("data_part/Disposable Personal Income.csv",
                 stringsAsFactors = FALSE)

## 30-year mortgage rate (weekly), MORTGAGE30US from FRED
mort <- read.csv("data_part/MORTGAGE30US.csv",
                 stringsAsFactors = FALSE)

## Rental index, monthly (CPI rent)
rent <- read.csv("data_part/Rental_Index.csv",
                 stringsAsFactors = FALSE)


## 1. Prepare a proper Date column in each table

## FHFA already has a 'date' column in "YYYY-MM-DD" format
fhfa$date <- as.Date(fhfa$date)

## DPI: use 'observation_date' as date
dpi$date <- as.Date(dpi$observation_date)

## Mortgage: weekly dates in 'observation_date'
mort$obs_date <- as.Date(mort$observation_date)

## Rental index: monthly dates in 'observation_date'
```

```r
rent$date <- as.Date(rent$observation_date)

## 2. Keep FHFA observations from year >= 2000

fhfa$year <- as.numeric(format(fhfa$date, "%Y"))
fhfa_2000 <- fhfa[fhfa$year >= 2000, ]

## 3. Construct monthly wage_income from DPI

## IMPORTANT:
## Column 'A229RC0' is the original DPI series in the CSV.
## Divided by 12 to approximate monthly disposable income.
dpi$wage_income <- dpi$A229RC0 / 12

## Keep only date and wage_income
dpi_sub <- dpi[, c("date", "wage_income")]

## 4. Aggregate weekly mortgage rates to monthly

## Extract year-month from weekly dates
mort$year  <- as.numeric(format(mort$obs_date, "%Y"))
mort$month <- as.numeric(format(mort$obs_date, "%m"))

## Create a "YYYY-MM" key
mort$ym <- paste(mort$year,
                 sprintf("%02d", mort$month),
                 sep = "-")

## Compute monthly average mortgage rate
mort_monthly <- aggregate(MORTGAGE30US ~ ym,
                          data = mort,
                          FUN = mean, na.rm = TRUE)

## Create a Date for each year-month (use 1st day of month)
mort_monthly$date <- as.Date(paste(mort_monthly$ym, "01", sep = "-"))

## Keep only date and monthly mortgage rate
mort_sub <- mort_monthly[, c("date", "MORTGAGE30US")]
names(mort_sub)[2] <- "mortgage_rate"

## 5. Prepare house_price and rent_cost

## FHFA: use index_sa as house_price
fhfa_sub <- fhfa_2000[, c("date", "index_sa", "year")]
names(fhfa_sub)[names(fhfa_sub) == "index_sa"] <- "house_price"

## Rental index: use CUUR0000SEHA as rent_cost
rent_sub <- rent[, c("date", "CUUR0000SEHA")]
names(rent_sub)[2] <- "rent_cost"

## 6. Merge all four datasets by 'date'

## Merge FHFA + DPI
df <- merge(fhfa_sub, dpi_sub, by = "date", all = FALSE)
```

```r
## Merge with mortgage (monthly average)
df <- merge(df, mort_sub, by = "date", all = FALSE)

## Merge with rental index
df <- merge(df, rent_sub, by = "date", all = FALSE)

## Recompute year from the final 'date' just to be safe
df$year <- as.numeric(format(df$date, "%Y"))

## 7. Construct affordability indices

df$home_aff <- df$house_price / df$wage_income
df$rent_aff <- df$rent_cost / df$wage_income

## 8. Define 'period' groups based on year

df$period <- NA_character_

df$period[df$year >= 2000 & df$year <= 2007] <- "2000-2007"
df$period[df$year >= 2008 & df$year <= 2013] <- "2008-2013"
df$period[df$year >= 2014 & df$year <= 2019] <- "2014-2019"
df$period[df$year >= 2020]                   <- "2020-2025"

## 9. High-affordability indicators
## home_aff_high = 1 if home_aff above median, 0 otherwise
## rent_aff_high = 1 if rent_aff above median, 0 otherwise

home_med <- median(df$home_aff, na.rm = TRUE)
rent_med <- median(df$rent_aff, na.rm = TRUE)

df$home_aff_high <- ifelse(df$home_aff > home_med, 1L, 0L)
df$rent_aff_high <- ifelse(df$rent_aff > rent_med, 1L, 0L)

## 10. Select final columns and sort by date

final_cols <- c("date",
                "house_price",
                "rent_cost",
                "wage_income",
                "mortgage_rate",
                "home_aff",
                "rent_aff",
                "period",
                "home_aff_high",
                "rent_aff_high",
                "year")

final_df <- df[, final_cols]

## Sort by date in ascending order
final_df <- final_df[order(final_df$date), ]

## 11. Export the final clean dataset
```

```r
write.csv(final_df,
          file = "data_part/housing_affordability_monthly_final.csv",
          row.names = FALSE)
```

2.2 States and Regions' Data

```r
## 0. Read all required CSV files

fhfa_us <- read.csv("data_part/fhfa_us_quarterly_sa.csv", stringsAsFactors = FALSE)
fhfa_hi <- read.csv("data_part/fhfa_hi_quarterly.csv", stringsAsFactors = FALSE)
fhfa_wv <- read.csv("data_part/fhfa_wv_quarterly.csv", stringsAsFactors = FALSE)
fhfa_ut <- read.csv("data_part/fhfa_ut_quarterly.csv", stringsAsFactors = FALSE)
fhfa_al <- read.csv("data_part/fhfa_al_quarterly.csv", stringsAsFactors = FALSE)
fhfa_dc <- read.csv("data_part/fhfa_dc_quarterly.csv", stringsAsFactors = FALSE)
fhfa_ms <- read.csv("data_part/fhfa_ms_quarterly.csv", stringsAsFactors = FALSE)

pi_annual <- read.csv("data_part/personal income.csv",
                      stringsAsFactors = FALSE,
                      check.names = FALSE)

dpi_q <- read.csv("data_part/Quarterly_Disposable personal income.csv",
                  stringsAsFactors = FALSE)

## 1. Prepare US Quarterly Income (divide by 12)

dpi_q$date <- as.Date(dpi_q$observation_date)
dpi_q$year <- as.numeric(format(dpi_q$date, "%Y"))
dpi_q$quarter <- as.numeric(format(dpi_q$date, "%m")) %/% 3 + 1

dpi_q$income_US <- dpi_q$A229RC0Q052SBEA / 12  ## per-quarter income

dpi_us_quarterly <- dpi_q[, c("year", "quarter", "income_US")]

## 2. Quarterly HPI helper function

make_q_hpi <- function(df, state_code) {
  tmp <- df[, c("yr", "period", "index_sa")]
  names(tmp) <- c("year", "quarter", "HPI_state")
  tmp$state_code <- state_code
  tmp
}

hpi_us_q <- fhfa_us[, c("yr", "period", "index_sa")]
names(hpi_us_q) <- c("year", "quarter", "HPI_US")

hpi_hi_q <- make_q_hpi(fhfa_hi, "HI")
hpi_wv_q <- make_q_hpi(fhfa_wv, "WV")
hpi_ut_q <- make_q_hpi(fhfa_ut, "UT")
hpi_al_q <- make_q_hpi(fhfa_al, "AL")
hpi_dc_q <- make_q_hpi(fhfa_dc, "DC")
```

```r
hpi_ms_q <- make_q_hpi(fhfa_ms, "MS")

## Combine state-level HPI
hpi_states_q <- rbind(hpi_hi_q, hpi_wv_q, hpi_ut_q,
                      hpi_al_q, hpi_dc_q, hpi_ms_q)

## 3. Convert annual state income into quarterly income

## Keep only US + our 6 states
pi_annual$state_code <- NA
pi_annual$state_code[pi_annual$GeoName == "United States *"]      <- "US"
pi_annual$state_code[pi_annual$GeoName == "Hawaii *"]             <- "HI"
pi_annual$state_code[pi_annual$GeoName == "West Virginia"]        <- "WV"
pi_annual$state_code[pi_annual$GeoName == "Utah"]                 <- "UT"
pi_annual$state_code[pi_annual$GeoName == "Alabama"]              <- "AL"
pi_annual$state_code[pi_annual$GeoName == "District of Columbia"] <- "DC"
pi_annual$state_code[pi_annual$GeoName == "Mississippi"]          <- "MS"

pi_annual <- pi_annual[!is.na(pi_annual$state_code), ]

years_vec <- 2000:2024
year_cols <- as.character(years_vec)

pi_long <- reshape(pi_annual,
                   varying = year_cols,
                   v.names = "income_annual",
                   timevar = "year",
                   times = years_vec,
                   idvar = "state_code",
                   direction = "long")

pi_long <- pi_long[, c("state_code", "year", "income_annual")]

## Expand to quarterly using linear interpolation
make_quarterly_income <- function(df_state) {
  y <- df_state$income_annual
  yrs <- df_state$year

  t <- yrs * 4  ## map each year to a time point (every 4 units)

  tq <- seq(min(t), max(t), by = 1)  ## quarterly time points

  yq <- approx(x = t, y = y, xout = tq)$y

  years_q <- tq %/% 4
  quarter_q <- tq %% 4 + 1

  data.frame(state_code = df_state$state_code[1],
             year = years_q,
             quarter = quarter_q,
             income_state = yq/12)
}

income_states_q <- do.call(rbind,
```

```
                          lapply(split(pi_long, pi_long$state_code),
                                 make_quarterly_income))

## US quarterly income already prepared in dpi_us_quarterly

## 4. Build state-level AffordGap (relative to US) and AffordLevel (relative to own base)

build_affordgap_q <- function(state_code) {

  # Subset state-level HPI (quarterly)
  hpi_s <- hpi_states_q[hpi_states_q$state_code == state_code,
                        c("year", "quarter", "HPI_state", "state_code")]

  # US HPI (quarterly)
  hpi_us <- hpi_us_q

  # State-level quarterly income (interpolated)
  inc_s <- income_states_q[income_states_q$state_code == state_code,
                           c("year", "quarter", "income_state")]

  # US quarterly income (from DPI)
  inc_us <- dpi_us_quarterly

  # Merge all components by year and quarter
  df <- merge(hpi_s, hpi_us, by = c("year", "quarter"), all = FALSE)
  df <- merge(df, inc_s, by = c("year", "quarter"), all = FALSE)
  df <- merge(df, inc_us, by = c("year", "quarter"), all = FALSE)

  # Ensure rows are ordered in time
  df <- df[order(df$year, df$quarter), ]

  ## Relative to US: AffordGap (your original measure)
  df$RelPrice  <- df$HPI_state    / df$HPI_US
  df$RelIncome <- df$income_state / df$income_US
  df$AffordGap <- df$RelPrice / df$RelIncome

  ## Relative to own base: AffordLevel_{s,t}
  ## Use the earliest available quarter for this state as the base (in your data: 2000
   ↪  Q1)
  base_hpi <- df$HPI_state[1L]
  base_inc <- df$income_state[1L]

  # State-specific price and income indices (base period = 1)
  df$PriceIndexSelf  <- df$HPI_state    / base_hpi
  df$IncomeIndexSelf <- df$income_state / base_inc

  # AffordLevel: change in affordability relative to the state's own base period
  df$AffordLevel <- df$PriceIndexSelf / df$IncomeIndexSelf

  # Make sure state_code column is present and correct
  df$state_code <- state_code

  df
}
```

```
## Build quarterly series for each selected state
ag_hi <- build_affordgap_q("HI")
ag_wv <- build_affordgap_q("WV")
ag_ut <- build_affordgap_q("UT")
ag_al <- build_affordgap_q("AL")
ag_dc <- build_affordgap_q("DC")
ag_ms <- build_affordgap_q("MS")

## Combine all states into one data frame
affordgap_quarterly <- rbind(ag_hi, ag_wv, ag_ut,
                             ag_al, ag_dc, ag_ms)

## Sort by state, year, and quarter
affordgap_quarterly <- affordgap_quarterly[
  order(affordgap_quarterly$state_code,
        affordgap_quarterly$year,
        affordgap_quarterly$quarter),
]

## Export to CSV
write.csv(affordgap_quarterly,
          "data_part/state_affordgap_quarterly.csv",
          row.names = FALSE)
```

3. EDA

3.1 Nation

```
if (!require("tidyverse")) install.packages("tidyverse")
library(tidyverse)
library(lubridate)

# 1. Load Data
df <- read_csv("data_part/housing_affordability_monthly_final.csv")
```

```
Rows: 308 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): period
dbl  (9): house_price, rent_cost, wage_income, mortgage_rate, home_aff, rent...
date (1): date

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
df <- df %>% mutate(date = ymd(date))

# 2. Define Structural Breaks Data Frame
# To ensure geom_vline generates a legend entry,
```

```r
# it is best to pass the breaks as a separate data frame
breaks_df <- data.frame(
  date = as.Date(c("2008-01-01", "2014-01-01", "2020-01-01")),
  type = "Structural Breaks"  # This string will appear in the legend
)

# 3. Plotting
p <- ggplot(df) +

  # Layer 1: Home Affordability (Blue)
  geom_line(
    aes(x = date, y = home_aff, color = "Home Affordability Index"),
    size = 1.2
  ) +

  # Layer 2: Rent Affordability (Orange)
  geom_line(
    aes(x = date, y = rent_aff, color = "Rent Affordability Index"),
    size = 1.2
  ) +

  # Layer 3: Structural Break Lines
  geom_vline(
    data = breaks_df,
    aes(xintercept = date, linetype = type),
    color = "red",
    size = 0.8
  ) +

  # Scale Formatting
  scale_color_manual(
    name = "Metric",
    values = c(
      "Home Affordability Index" = "#1f77b4",
      "Rent Affordability Index" = "#ff7f0e"
    )
  ) +

  scale_linetype_manual(
    name = "Event",
    values = c("Structural Breaks" = "dashed")
  ) +

  # Labels
  labs(
    title = "Structural Breaks in Affordability Indices",
    x = "Year",
    y = "Affordability Index"
  ) +

  scale_x_date(date_breaks = "2 years", date_labels = "%Y") +

  # Theme Adjustments (Optimized for Spacing)
  theme_minimal() +
```

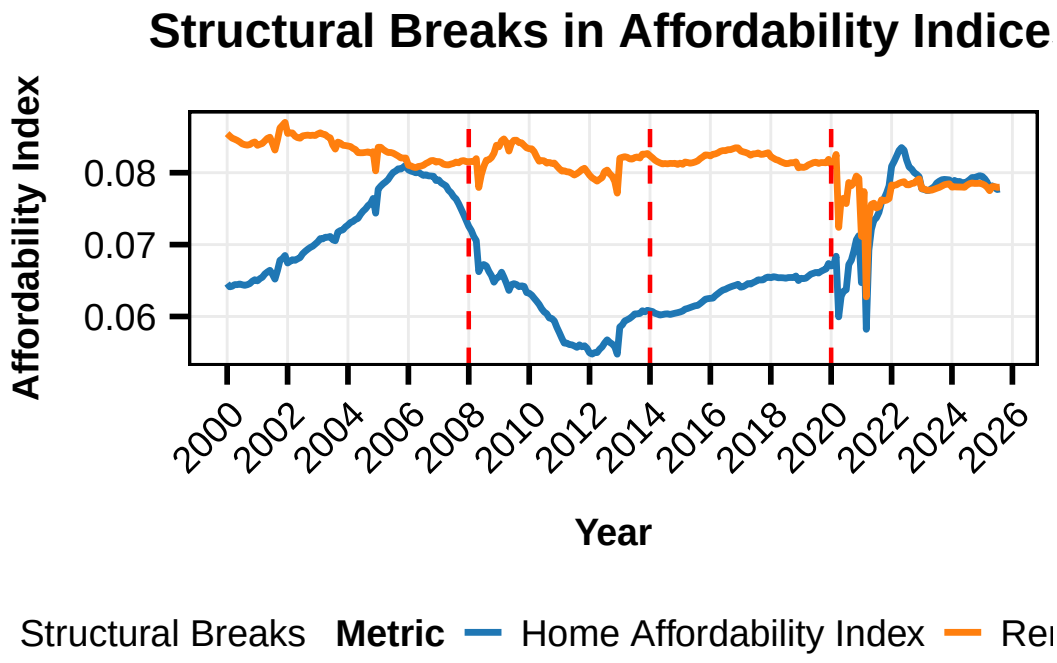
```
  theme(
    # 1. Title styling
    plot.title = element_text(
      face = "bold",
      size = 18,
      margin = margin(b = 20),
      hjust = 0.5
    ),

    # 2. Axis titles
    axis.title = element_text(face = "bold", size = 14),
    axis.title.x = element_text(margin = margin(t = 14)),
    axis.title.y = element_text(margin = margin(r = 14)),

    # 3. Axis tick labels
    axis.text = element_text(size = 14, color = "black"),
    axis.text.x = element_text(angle = 45, hjust = 1),

    # 4. Axis tick marks
    axis.ticks = element_line(color = "black", size = 1),
    axis.ticks.length = unit(0.25, "cm"),

    # 5. Legend settings
    legend.position = "bottom",
    legend.box = "horizontal",
    legend.title = element_text(face = "bold", size = 14),
    legend.text = element_text(size = 14),
    legend.margin = margin(t = 10),

    # Panel and Grid
    panel.grid.minor = element_blank(),
    panel.border = element_rect(color = "black", fill = NA, size = 1),


  )
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.
i Please use the `linewidth` argument instead.
```

```
Warning: The `size` argument of `element_rect()` is deprecated as of ggplot2 3.4.0.
i Please use the `linewidth` argument instead.
```

```
print(p)
```



3.2 States and Regions

```
# Regional Affordability (Full State Names Version - Including DC)


if (!require("tidyverse")) install.packages("tidyverse")
library(tidyverse)
library(lubridate)

# Load Data
df <- read_csv("data_part/state_affordgap_quarterly.csv")
```

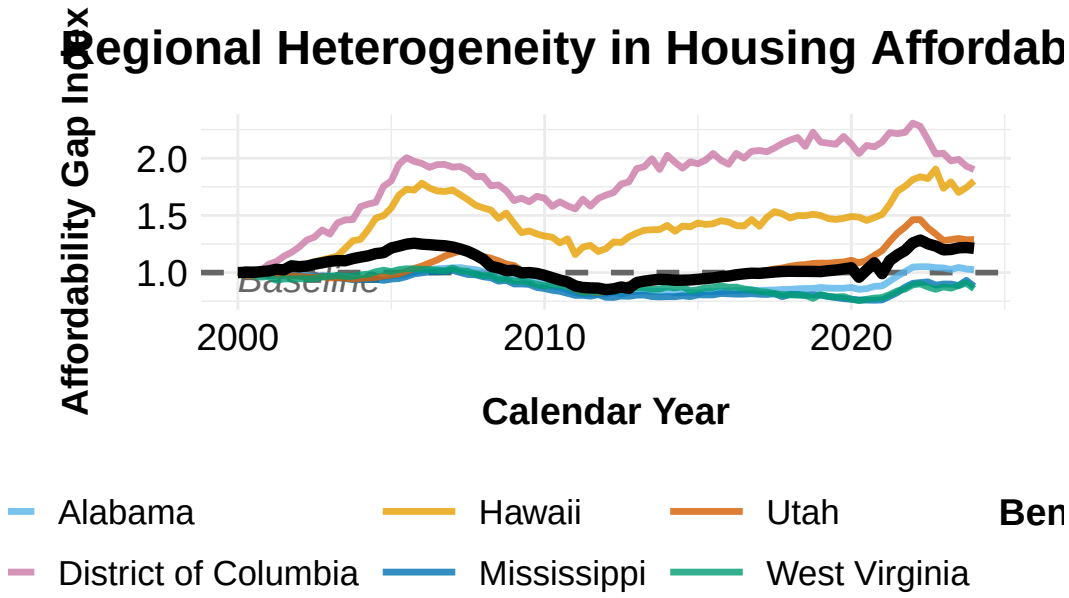
```
Rows: 1164 Columns: 13
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): state_code
dbl (12): year, quarter, HPI_state, HPI_US, income_state, income_US, RelPric...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Prepare National Data (Baseline)
df_us <- df %>%
  select(year, quarter, HPI_US, income_US) %>%
  distinct() %>%
  na.omit() %>%
  group_by(year, quarter) %>%
  summarise(HPI_US = mean(HPI_US, na.rm = TRUE),
            income_US = mean(income_US, na.rm = TRUE),
```

```
            .groups = 'drop') %>%
  arrange(year, quarter)

# Calculate Index (Base 2000 Q1 = 1.0)
base_hpi <- df_us$HPI_US[1]
base_inc <- df_us$income_US[1]

df_us <- df_us %>%
  mutate(
    Date = ymd(paste(year, (quarter*3)-2, "01", sep="-")),
    PriceIndex = HPI_US / base_hpi,
    IncomeIndex = income_US / base_inc,
    AffordLevel = PriceIndex / IncomeIndex,
    Entity = "National Average"
  )

# Prepare State Data (WITH FULL NAMES)
state_lookup <- c(
  "UT" = "Utah",
  "HI" = "Hawaii",
  "AL" = "Alabama",
  "MS" = "Mississippi",
  "WV" = "West Virginia",
  "DC" = "District of Columbia"
)

target_states <- names(state_lookup)

# Prepare State Data (df_states)
df_states <- df %>%
  filter(state_code %in% target_states) %>%
  select(year, quarter, state_code, HPI_state, income_state) %>%
  drop_na(HPI_state, income_state) %>%
  group_by(year, quarter, state_code) %>%
  summarise(
    HPI_state = mean(HPI_state, na.rm = TRUE),
    income_state = mean(income_state, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(state_code, year, quarter) %>%
  group_by(state_code) %>%
  mutate(
    Date = ymd(paste(year, (quarter * 3) - 2, "01", sep = "-")),
    PriceIndex = HPI_state / first(HPI_state),
    IncomeIndex = income_state / first(income_state),
    AffordLevel = PriceIndex / IncomeIndex,
    Entity = unname(state_lookup[state_code])
  ) %>%
  ungroup()

# Added Color for District of Columbia
my_colors <- c(
  "Utah"                 = "#D55E00",
  "Hawaii"               = "#E69F00",
```

```
  "Alabama"              = "#56B4E9",
  "Mississippi"          = "#0072B2",
  "West Virginia"        = "#009E73",
  "District of Columbia" = "#CC79A7"  # New Color (Reddish Purple)
)
p <- ggplot() +
  # Reference Line
  geom_hline(yintercept = 1, linetype = "dashed", color = "gray40", size = 1) +
  annotate("text", x = min(df_us$Date), y = 0.95,
           label = "Baseline",
           hjust = 0, size = 5, color = "gray40", fontface = "italic") +

  # Layer 1: States
  geom_line(data = df_states, aes(x = Date, y = AffordLevel, color = Entity),
            size = 1.2, alpha = 0.8) +

  # Layer 2: National Average
  geom_line(data = df_us, aes(x = Date, y = AffordLevel, linetype = "National Average"),
            color = "black", size = 2) +

  # Labels
  labs(
    title = "Regional Heterogeneity in Housing Affordability",
    y = "Affordability Gap Index",
    x = "Calendar Year",
    color = "Selected Regions",
    linetype = "Benchmark"
  ) +

  scale_color_manual(values = my_colors) +
  scale_linetype_manual(values = c("National Average" = "solid")) +

  # Theme Adjustments
  theme_minimal() +
  theme(
    plot.title = element_text(face = "bold", size = 18, margin = margin(b = 15), hjust =
    ↪  0.5),

    axis.title.y = element_text(face = "bold", size = 14, margin = margin(r = 15)),
    axis.title.x = element_text(face = "bold", size = 14, margin = margin(t = 15)),
    axis.text = element_text(size = 14, color = "black"),

    legend.position = "bottom",
    legend.box = "horizontal",

    legend.title = element_text(face = "bold", size = 14),
    legend.text = element_text(size = 13),
    legend.key.width = unit(1.2, "cm"),

    legend.margin = margin(t = 10),

    plot.margin = margin(t = 20, r = 20, b = 20, l = 20, unit = "pt")
  )
```

```
print(p)
```



4. Tests

4.1 t-tests

```
# Load Data
df <- read_csv("data_part/housing_affordability_monthly_final.csv")
```

```
Rows: 308 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): period
dbl  (9): house_price, rent_cost, wage_income, mortgage_rate, home_aff, rent...
date (1): date

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
library(dplyr)
df_group <- df %>% group_by(period) %>%
    summarise(home_aff = mean(home_aff),
              rent_aff = mean(rent_aff))
print(df_group)
```

```
# A tibble: 4 x 3
  period    home_aff rent_aff
```

```
  <chr>        <dbl>    <dbl>
1 2000-2007   0.0726   0.0834
2 2008-2013   0.0609   0.0815
3 2014-2019   0.0636   0.0819
4 2020-2025   0.0758   0.0775
```

```r
# HomeAff: 2000-2007 vs 2020-2025
early  <- df$home_aff[df$period=="2000-2007"]
recent <- df$home_aff[df$period=="2020-2025"]
t.test(early, recent, var.equal=FALSE)
```

```
    Welch Two Sample t-test

data:  early and recent
t = -3.4392, df = 139.98, p-value = 0.0007689
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.004972857 -0.001342464
sample estimates:
 mean of x  mean of y
0.07263286 0.07579052
```

```r
# RentAff: 2013-2019 vs 2020-2025
mid    <- df$rent_aff[df$period=="2014-2019"]
recent <- df$rent_aff[df$period=="2020-2025"]
t.test(mid, recent, var.equal=FALSE)
```

```
    Welch Two Sample t-test

data:  mid and recent
t = 13.979, df = 78.481, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.003783623 0.005040120
sample estimates:
 mean of x  mean of y
0.08194478 0.07753291
```

```r
# ANOVA + Tukey HSD
anova_res <- aov(home_aff ~ period, data = df)
summary(anova_res)
```

```
             Df   Sum Sq  Mean Sq F value Pr(>F)
period        3 0.011148 0.003716   157.5 <2e-16 ***
Residuals   304 0.007173 0.000024
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
TukeyHSD(anova_res)
```

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = home_aff ~ period, data = df)

$period
                             diff           lwr          upr     p adj
2008-2013-2000-2007 -0.011751695 -0.0137079742 -0.009795416 0.0000000
2014-2019-2000-2007 -0.009013251 -0.0109695298 -0.007056971 0.0000000
2020-2025-2000-2007  0.003157661  0.0011687744  0.005146547 0.0003067
2014-2019-2008-2013  0.002738444  0.0006470939  0.004829795 0.0044905
2020-2025-2008-2013  0.014909355  0.0127874727  0.017031238 0.0000000
2020-2025-2014-2019  0.012170911  0.0100490283  0.014292794 0.0000000
```

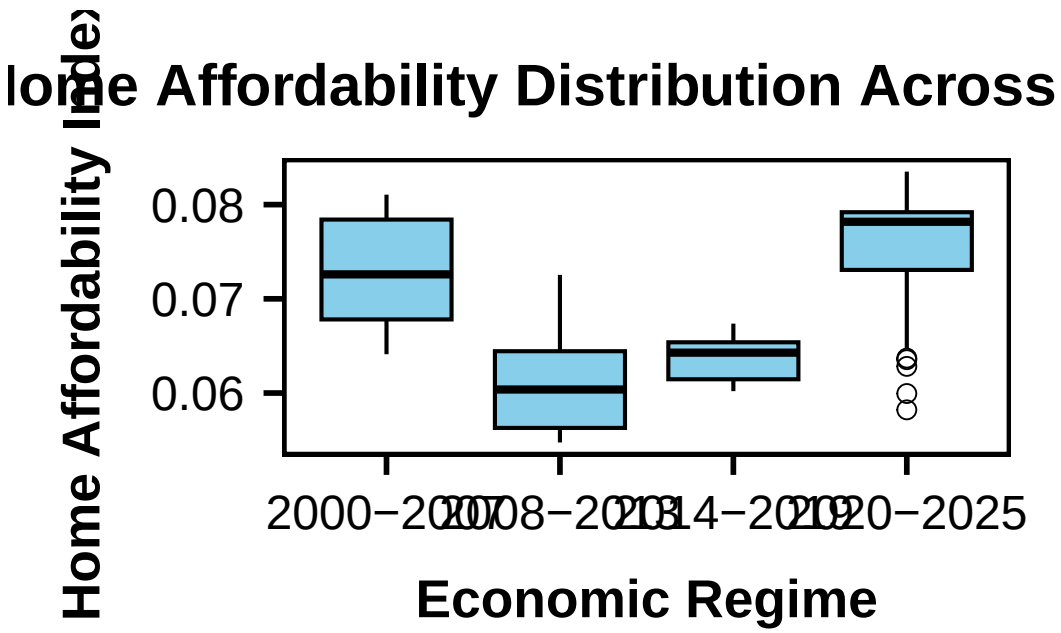
```r
if (!require("tidyverse")) install.packages("tidyverse")
library(tidyverse)

# 1. Load Data
df <- read_csv("data_part/housing_affordability_monthly_final.csv")
```

```
Rows: 308 Columns: 11
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): period
dbl  (9): house_price, rent_cost, wage_income, mortgage_rate, home_aff, rent...
date (1): date

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# 2. Reorder 'period' to ensure chronological order on the x-axis
# This step is critical; otherwise, R may sort factor levels alphabetically,
# placing "2000-2007" after "2020-2025"
# Adjust the level order below to match the actual labels in your data
df$period <- factor(
  df$period,
  levels = c("2000-2007", "2008-2013", "2014-2019", "2020-2025")
)

# 3. Plotting
p_box <- ggplot(df, aes(x = period, y = home_aff)) +

  # Boxplot Layer
  # fill = "skyblue": light blue fill
  # color = "black": black box outline for better contrast
  geom_boxplot(
    fill = "skyblue",
    color = "black",
    outlier.shape = 1,
    outlier.size = 3
```

```
  ) +

  # Labels
  labs(
    title = "Home Affordability Distribution Across Regimes",
    x = "Economic Regime",   # Improved x-axis label
    y = "Home Affordability Index"
  ) +

  # Theme Adjustments (Matching Previous Style)
  theme_minimal(base_size = 16) +
  theme(
    # 1. Centered and enlarged title
    plot.title = element_text(
      face = "bold",
      size = 22,
      margin = margin(b = 15),
      hjust = 0.5
    ),

    # 2. Enlarged axis titles
    axis.title = element_text(face = "bold", size = 20),
    axis.title.x = element_text(margin = margin(t = 15)),
    axis.title.y = element_text(margin = margin(r = 15)),

    # 3. Enlarged axis tick labels
    axis.text = element_text(size = 18, color = "black"),

    # Slight rotation of x-axis labels to prevent overlap
    # Optional: remove angle = 15 if labels are short
    axis.text.x = element_text(angle = 0, hjust = 0.5),

    # 4. Axis tick marks
    axis.ticks = element_line(color = "black", size = 1),
    axis.ticks.length = unit(0.25, "cm"),

    # 5. Panel styling: no grid lines + black border
    panel.grid = element_blank(),
    panel.border = element_rect(color = "black", fill = NA, size = 1.5),

    # 6. Plot margins
    plot.margin = margin(20, 20, 20, 20)
  )

# 4. Save Output
# Note: the object name is p_box, and the file name reflects the boxplot content
ggsave(
  "HomeAff_Boxplot_Regimes_Styled.png",
  plot = p_box,
  width = 14,
  height = 9,
  dpi = 300
)
```

```
print(p_box)
```



4.2 Chi-square test

```
rate_median <- median(df$mortgage_rate)
df$rate_high <- ifelse(df$mortgage_rate > rate_median, 1, 0)

cont_table <- table(df$rate_high, df$home_aff_high)
cont_table
```

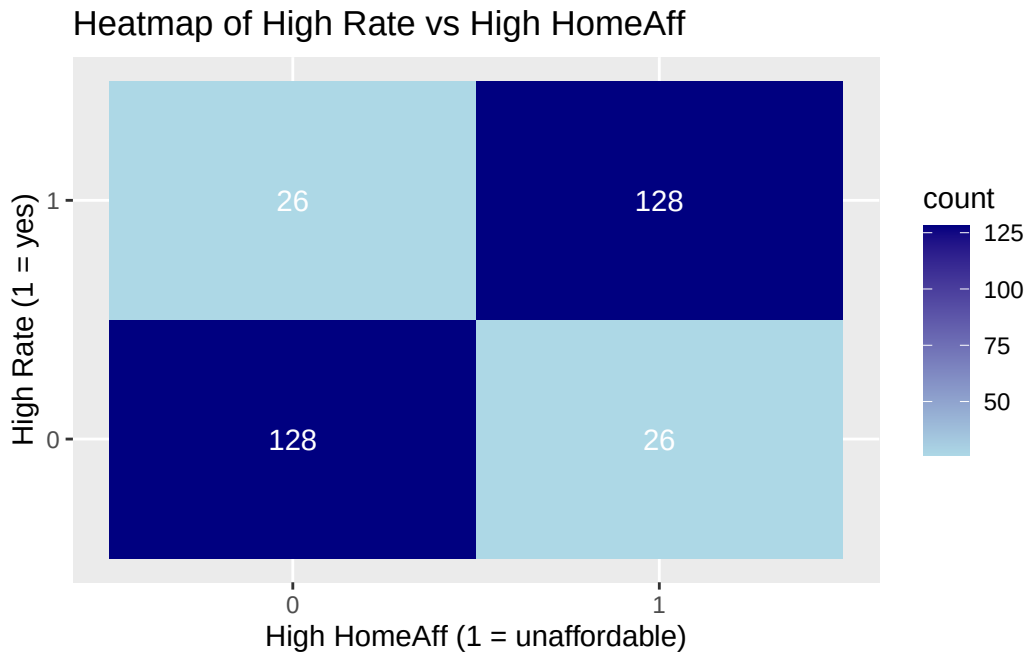
```
     0   1
  0 128  26
  1  26 128
```

```
chisq.test(cont_table)
```

```
    Pearson's Chi-squared test with Yates' continuity correction

data:  cont_table
X-squared = 132.48, df = 1, p-value < 2.2e-16
```

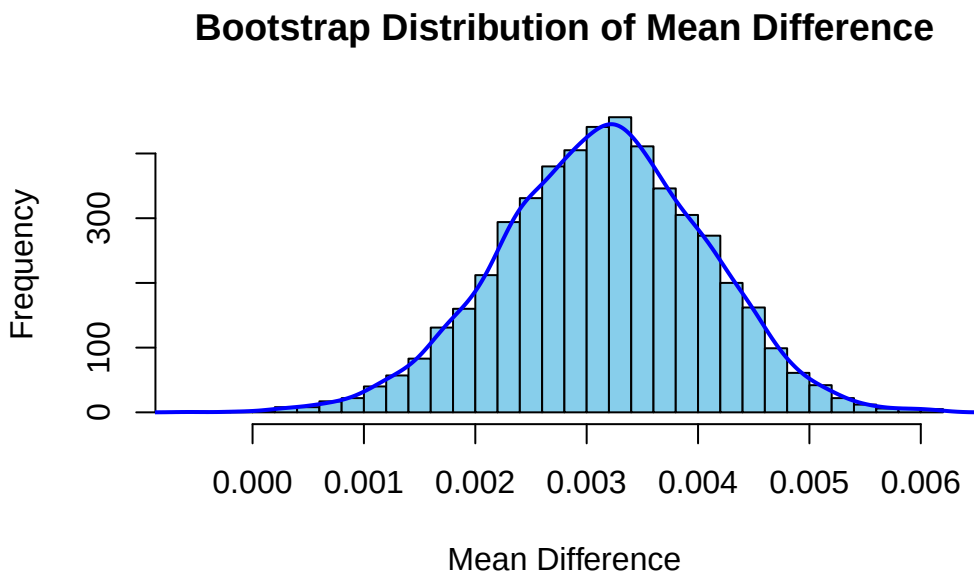
```
df_heat <- as.data.frame(cont_table)
colnames(df_heat) <- c("rate_high","home_aff_high","count")

ggplot(df_heat, aes(x=home_aff_high, y=rate_high, fill=count)) +
```

```
    geom_tile() +
    geom_text(aes(label=count), color="white") +
    scale_fill_gradient(low="lightblue", high="navy") +
    labs(title="Heatmap of High Rate vs High HomeAff",
         x="High HomeAff (1 = unaffordable)",
         y="High Rate (1 = yes)")
```



4.3 Bootstrap (Mean Difference & Correlation)

```
# Mean Diff
set.seed(123)

early  <- df$home_aff[df$period=="2000-2007"]
recent <- df$home_aff[df$period=="2020-2025"]

B <- 5000
boot_diff <- numeric(B)

for (i in 1:B) {
    boot_diff[i] <- mean(sample(recent, replace=TRUE)) -
                    mean(sample(early, replace=TRUE))
}

hist(boot_diff, breaks=40, col="skyblue",
     main="Bootstrap Distribution of Mean Difference",
     xlab="Mean Difference")
lines(density(boot_diff), col="blue", lwd=2)
```



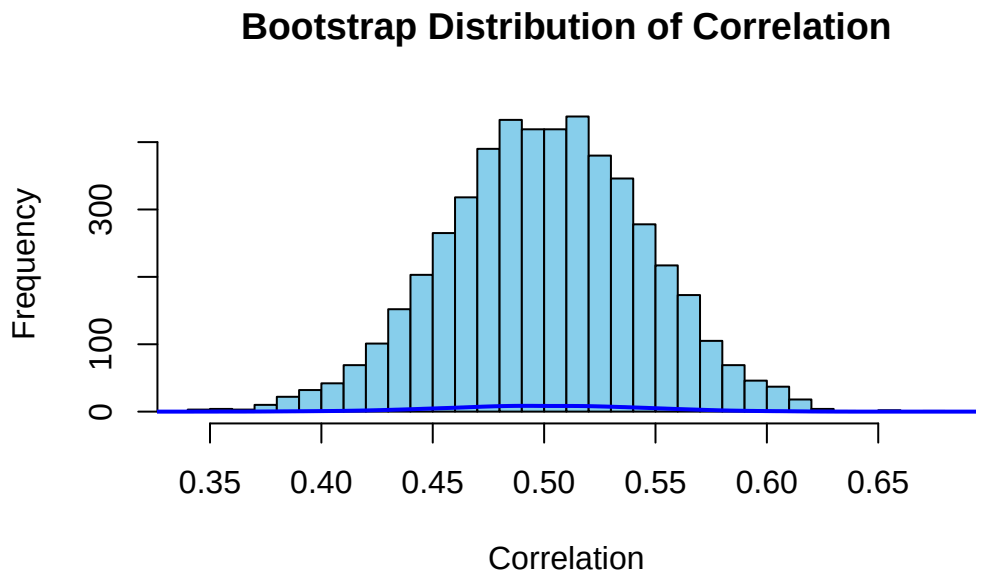

```
mean_diff <- mean(boot_diff)
ci <- quantile(boot_diff, c(0.025, 0.975))
se <- sd(boot_diff)
bias <- mean_diff - (mean(recent)-mean(early))

list(mean_diff=mean_diff, CI=ci, SE=se, bias=bias)
```

```
$mean_diff
[1] 0.003141182

$CI
       2.5%       97.5%
0.001289509 0.004875925

$SE
[1] 0.0009170381

$bias
[1] -1.647843e-05
```

4.4 Bootstrap Correlation (Mortgage vs HomeAff)

```
set.seed(123)

B <- 5000
boot_corr <- numeric(B)

for (i in 1:B) {
    idx <- sample(1:nrow(df), replace=TRUE)
```

```
    boot_corr[i] <- cor(df$mortgage_rate[idx], df$home_aff[idx])
}

hist(boot_corr, breaks=40, col="skyblue",
     main="Bootstrap Distribution of Correlation",
     xlab="Correlation")
lines(density(boot_corr), col="blue", lwd=2)
```




```
list(
    original = cor(df$mortgage_rate, df$home_aff),
    mean_corr = mean(boot_corr),
    CI = quantile(boot_corr, c(0.025, 0.975)),
    SE = sd(boot_corr),
    bias = mean(boot_corr) - cor(df$mortgage_rate, df$home_aff)
)
```

```
$original
[1] 0.5001435

$mean_corr
[1] 0.5006174

$CI
     2.5%     97.5%
0.4116102 0.5874951

$SE
[1] 0.04468781

$bias
[1] 0.0004738286
```

5. ARIMA Process

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
library(tidyverse)
library(lubridate)
library(forecast)
```

```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

```
library(tseries)
library(gridExtra)
library(knitr)
library(kableExtra)
```

```

Attaching package: 'kableExtra'

The following object is masked from 'package:dplyr':

    group_rows
```

```
library(astsa)
```

```

Attaching package: 'astsa'

The following object is masked from 'package:forecast':

    gas
```

```
library(ggplot2)
library(gridExtra)
library(lubridate)
library(zoo)
```

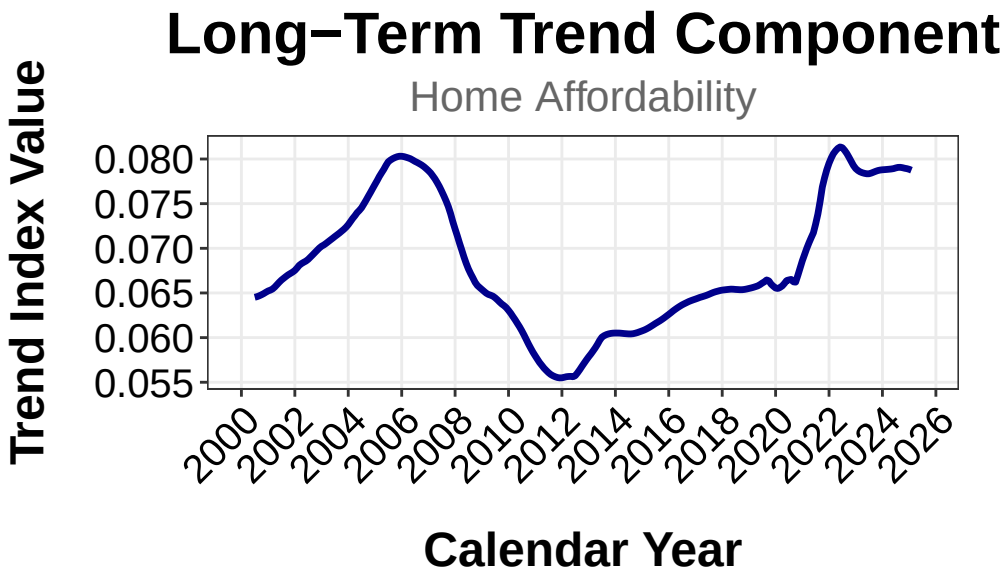
```

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
# 1. Data Loading & Preprocessing
df <- read.csv("data_part/housing_affordability_monthly_final.csv")

# Date conversion
```

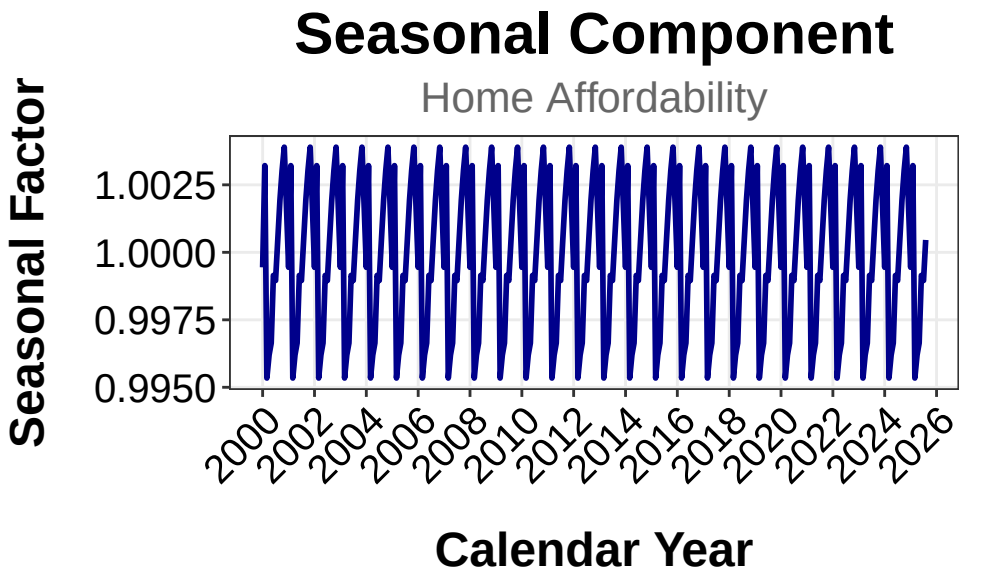
```r
df$date <- as.Date(paste0(df$date, "-01"))

# Create Time Series Objects (Frequency = 12 for monthly data)
# Start date extraction
start_year <- year(min(df$date))
start_month <- month(min(df$date))

ts_home <- ts(df$home_aff, start = c(start_year, start_month), frequency = 12)
ts_rent <- ts(df$rent_aff, start = c(start_year, start_month), frequency = 12)

# Remove NA if any
ts_home <- na.omit(ts_home)
ts_rent <- na.omit(ts_rent)
```

```r
# 1. Update Theme Settings (with 45-degree rotated x-axis labels)
theme_large <- theme_bw() +
  theme(
    # Plot title settings
    plot.title = element_text(size = 22, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 16, color = "gray40", hjust = 0.5),

    # Axis title settings
    axis.title.x = element_text(size = 18, face = "bold", margin = margin(t = 15)),
    axis.title.y = element_text(size = 18, face = "bold", margin = margin(r = 15)),

    # Axis tick labels (general settings)
    axis.text = element_text(size = 15, color = "black"),

    # Key modification: rotate x-axis tick labels by 45 degrees
    # angle = 45: rotate labels by 45 degrees
    # hjust = 1: right-align labels (critical to align text with tick marks)
    axis.text.x = element_text(angle = 45, hjust = 1),

    # Plot margins (extra bottom margin for rotated labels)
    plot.margin = margin(20, 20, 20, 20),

    # Remove minor grid lines
    panel.grid.minor = element_blank()
  )

# Generate monthly date sequence matching the time series length
dates_seq <- seq(
  from = as.Date(paste0(start(ts_home)[1], "-", start(ts_home)[2], "-01")),
  by = "month",
  length.out = length(ts_home)
)

# Perform multiplicative time series decomposition
decomp_home <- decompose(ts_home, type = "multiplicative")

# Convert decomposition results into a data frame
df_decomp <- data.frame(
  Date = dates_seq,
  Trend = as.numeric(decomp_home$trend),
```

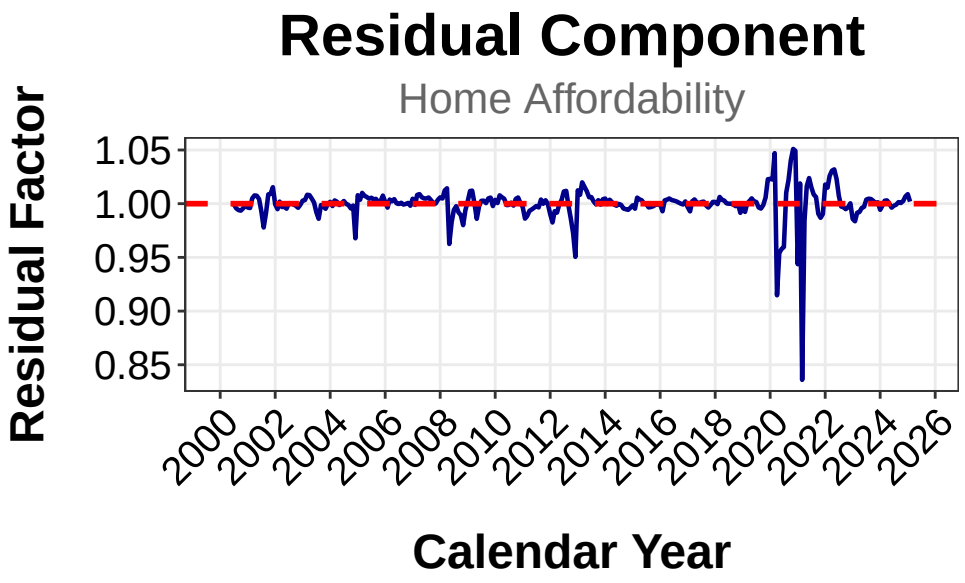
```
  Seasonal = as.numeric(decomp_home$seasonal),
  Random = as.numeric(decomp_home$random)
)

# Plot 1: Trend Component
p_trend <- ggplot(df_decomp, aes(x = Date, y = Trend)) +
  geom_line(color = "darkblue", size = 1.2) +
  # Slightly denser x-axis ticks; rotated labels allow more room
  scale_x_date(breaks = date_breaks("2 years"), labels = date_format("%Y")) +
  labs(
    title = "Long-Term Trend Component",
    subtitle = "Home Affordability",
    x = "Calendar Year",
    y = "Trend Index Value"
  ) +
  theme_large

# Plot 2: Seasonal Component
p_seasonal <- ggplot(df_decomp, aes(x = Date, y = Seasonal)) +
  geom_line(color = "darkblue", size = 1) +
  scale_x_date(breaks = date_breaks("2 years"), labels = date_format("%Y")) +
  labs(
    title = "Seasonal Component",
    subtitle = "Home Affordability",
    x = "Calendar Year",
    y = "Seasonal Factor"
  ) +
  theme_large

# Plot 3: Remainder / Residual Component
p_random <- ggplot(df_decomp, aes(x = Date, y = Random)) +
  geom_line(color = "darkblue", size = 0.8) +
  geom_hline(yintercept = 1, linetype = "dashed", color = "red", size = 1) +
  scale_x_date(breaks = date_breaks("2 years"), labels = date_format("%Y")) +
  labs(
    title = "Residual Component",
    subtitle = "Home Affordability",
    x = "Calendar Year",
    y = "Residual Factor"
  ) +
  theme_large

# Output plots
print(p_trend)
```




```
print(p_seasonal)
```



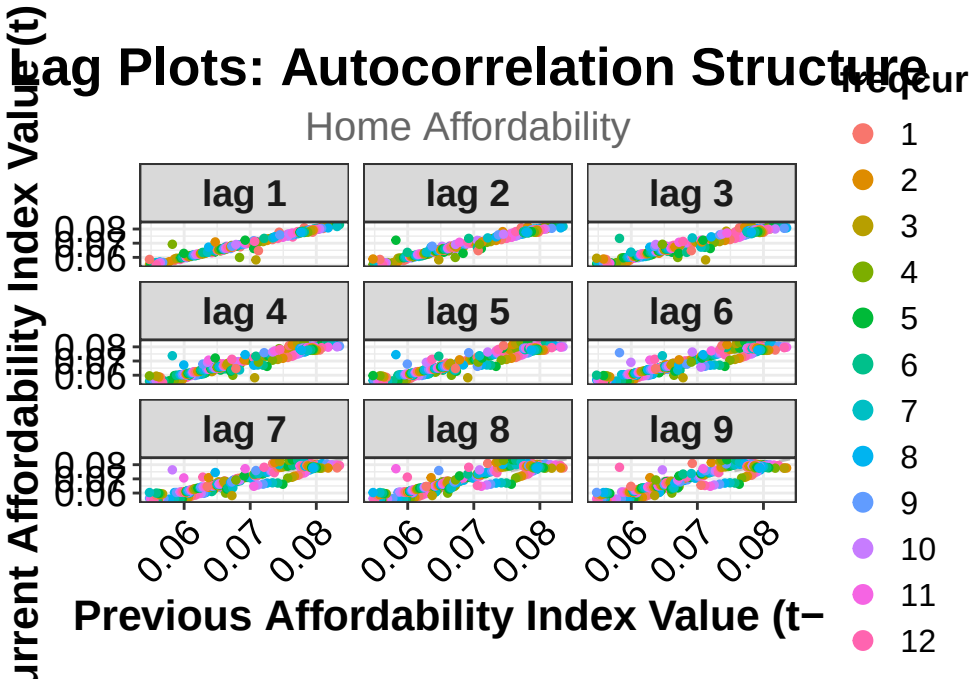

```
print(p_random)
```



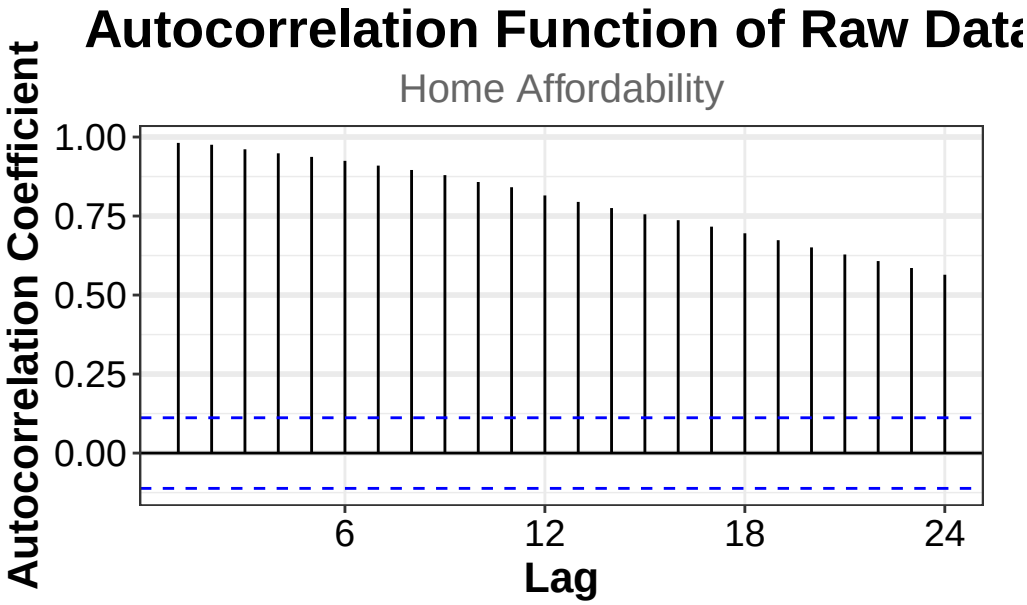

```
# Save plots to files
ggsave("Decomposition_1_Trend.png", plot = p_trend, width = 8, height = 5, dpi = 300)
ggsave("Decomposition_2_Seasonal.png", plot = p_seasonal, width = 8, height = 5, dpi =
 ↪  300)
ggsave("Decomposition_3_Residual.png", plot = p_random, width = 8, height = 5, dpi = 300)
```

```
library(ggplot2)
library(forecast)

# Define Large-Font Theme (unchanged)
theme_large_plots <- theme_bw() +
  theme(
    plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 15, color = "gray40", hjust = 0.5),
    axis.title = element_text(size = 16, face = "bold"),
    axis.text = element_text(size = 14, color = "black"),
    strip.text = element_text(size = 14, face = "bold"),
    plot.margin = margin(15, 15, 15, 15)
  )

# 1. Corrected Lag Plot (restore color encoding)
# Key point: removed geom_point(color = "darkblue")
# gglagplot automatically colors points by month (default rainbow palette)

p_lag <- gglagplot(ts_home, lags = 9, do.lines = FALSE) +

  # Descriptive labels and units
  labs(
    title = "Lag Plots: Autocorrelation Structure",
    subtitle = "Home Affordability",
    x = "Previous Affordability Index Value (t-k)",
    y = "Current Affordability Index Value (t)"
```

```
  ) +

  # Apply large-font theme
  theme_large_plots +

  # Prevent x-axis label overlap and improve legend readability
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),

    # Increase legend text size for clearer color distinction
    legend.text = element_text(size = 12),
    legend.title = element_text(size = 14, face = "bold")
  ) +

  # Increase legend point size for better visibility (optional)
  # This overrides the legend aesthetics without changing the plot itself
  guides(colour = guide_legend(override.aes = list(size = 3)))

# 2. ACF Plot (unchanged)
p_acf <- ggAcf(ts_home, lag.max = 24) +
  labs(
    title = "Autocorrelation Function of Raw Data",
    subtitle = "Home Affordability",
    x = "Lag",
    y = "Autocorrelation Coefficient"
  ) +
  theme_large_plots +
  theme(
    panel.grid.major.y = element_line(size = 1)
  )

# Output
print(p_lag)
```

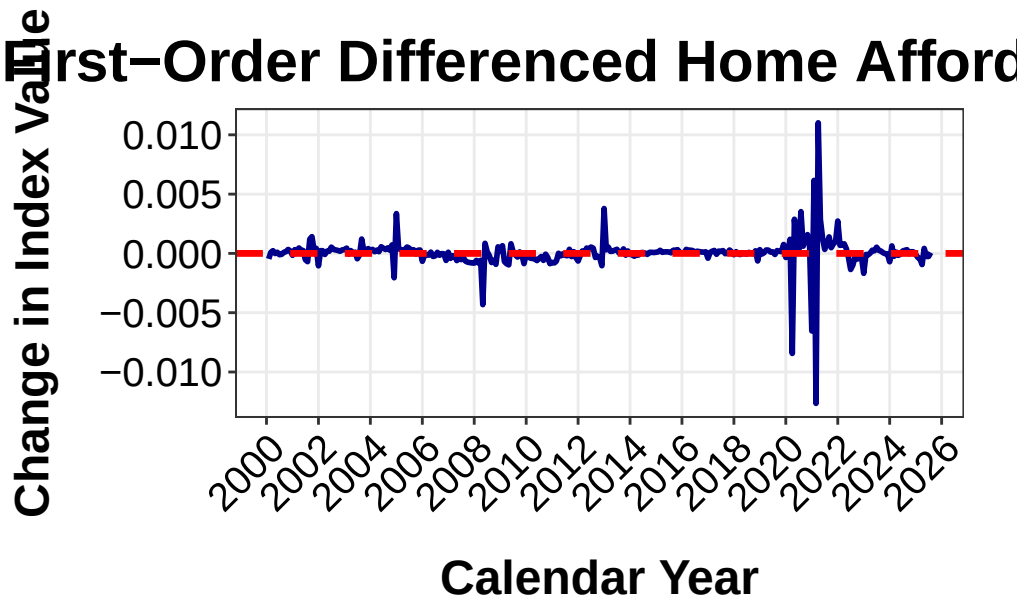



```
print(p_acf)
```




```
# Augmented
dates_diff <- df$date[-1]

ts_home_diff <- diff(ts_home, differences = 1)

df_diff_plot <- data.frame(
  Date = dates_diff,
```

```r
  Diff_Value = as.numeric(ts_home_diff)
)

# 2. Define Large-Font Theme (Large Fonts Theme)
# Includes 45-degree rotated x-axis labels to prevent overlap
theme_large_diff <- theme_bw() +
  theme(
    # Plot titles (size 20+, bold)
    plot.title = element_text(size = 22, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 16, color = "gray40", hjust = 0.5),

    # Axis titles (size 18+)
    axis.title.x = element_text(size = 18, face = "bold", margin = margin(t = 15)),
    axis.title.y = element_text(size = 18, face = "bold", margin = margin(r = 15)),

    # Axis tick labels (size 15+, rotated to avoid overlap)
    axis.text.x = element_text(size = 15, color = "black", angle = 45, hjust = 1),
    axis.text.y = element_text(size = 15, color = "black"),

    # Plot margins
    plot.margin = margin(20, 20, 20, 20),

    # Remove minor grid lines
    panel.grid.minor = element_blank()
  )

# 3. Plot First-Order Differenced Series
p_diff <- ggplot(df_diff_plot, aes(x = Date, y = Diff_Value)) +

  # B. Differenced time series line (distinct color for clarity)
  geom_line(color = "darkblue", size = 1) +

  # A. Zero reference line (critical for assessing stationarity)
  geom_hline(yintercept = 0, linetype = "dashed", color = "red", size = 1.2) +

  # C. Time axis formatting
  scale_x_date(breaks = date_breaks("2 years"), labels = date_format("%Y")) +

  # D. Descriptive labels and units
  labs(
    title = "First-Order Differenced Home Affordability",
    x = "Calendar Year",
    y = "Change in Index Value"
  ) +

  # E. Apply large-font theme
  theme_large_diff

# Output
print(p_diff)
```

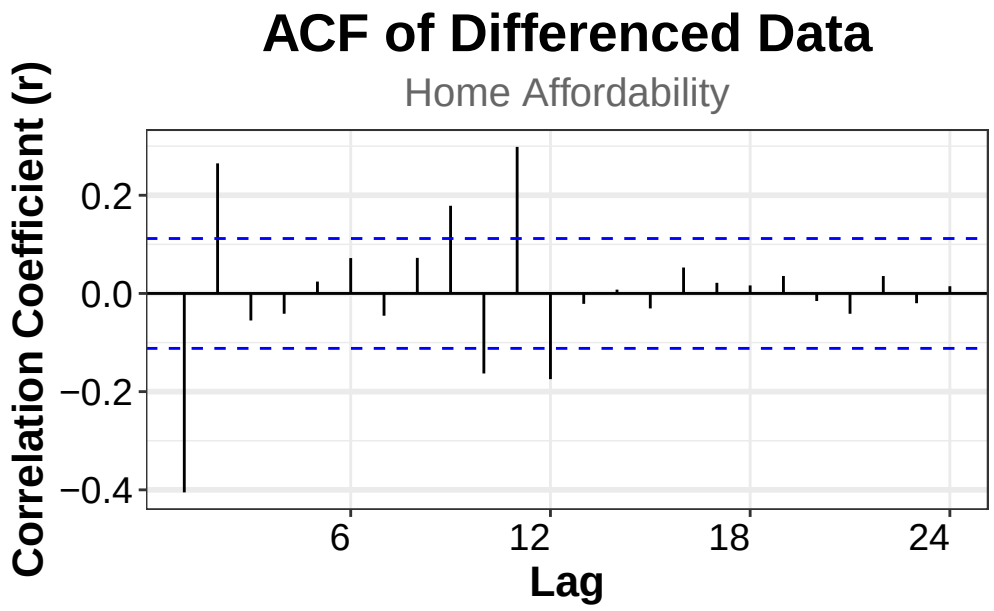



```
# Augmented Dickey-Fuller (ADF) test on differenced series
print(adf.test(ts_home_diff))
```

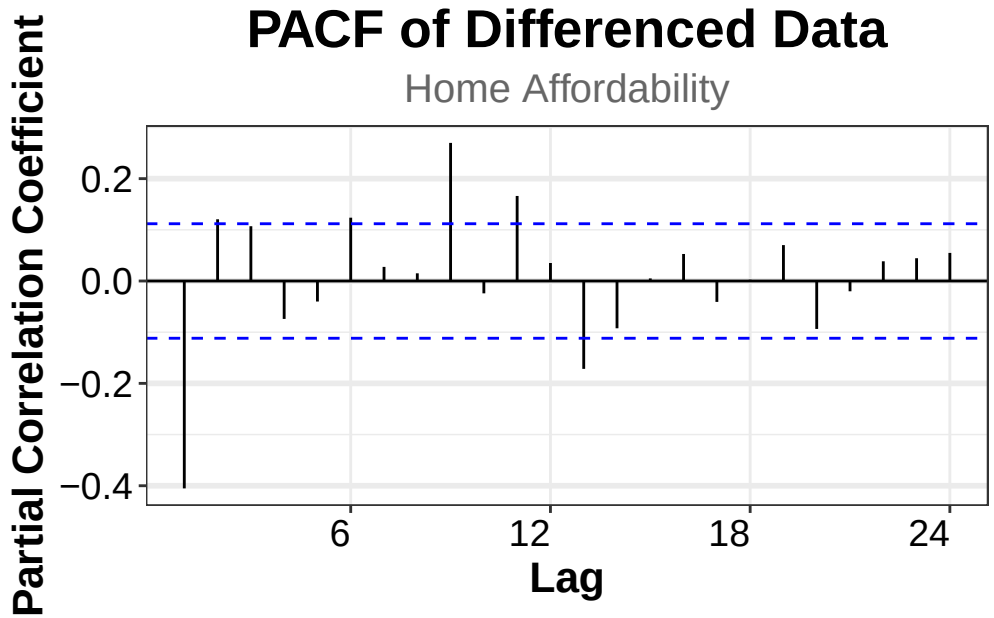
```
    Augmented Dickey-Fuller Test

data:  ts_home_diff
Dickey-Fuller = -5.8419, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

```
library(ggplot2)
library(forecast)

# 1. Define Large-Font Theme (Large Fonts Theme)
theme_large_acf <- theme_bw() +
  theme(
    # Plot titles (size 20+, bold)
    plot.title = element_text(size = 20, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 15, color = "gray40", hjust = 0.5),

    # Axis titles (size 16+)
    axis.title = element_text(size = 16, face = "bold"),

    # Axis tick labels (size 14+, prevent overlap)
    # Even simple numeric labels (1, 2, 3, …) may overlap at large font sizes,
    # so a 45-degree rotation is retained for safety
    axis.text.x = element_text(size = 14, color = "black", hjust = 1),
    axis.text.y = element_text(size = 14, color = "black"),

    # Emphasize horizontal reference grid lines (y = 0)
    panel.grid.major.y = element_line(size = 1),
```

```
    # Plot margins
    plot.margin = margin(15, 15, 15, 15)
  )

# 2. Plot ACF (used to identify MA terms - q)
p_acf_diff <- ggAcf(ts_home_diff, lag.max = 24) +

  # Descriptive labels and units
  labs(
    title = "ACF of Differenced Data",
    subtitle = "Home Affordability",
    x = "Lag",
    y = "Correlation Coefficient (r)"
  ) +

  # Apply large-font theme
  theme_large_acf

# 3. Plot PACF (used to identify AR terms - p)
p_pacf_diff <- ggPacf(ts_home_diff, lag.max = 24) +

  # Descriptive labels and units
  labs(
    title = "PACF of Differenced Data",
    subtitle = "Home Affordability",
    x = "Lag",
    y = "Partial Correlation Coefficient"
  ) +

  # Apply large-font theme
  theme_large_acf

# 4. Output
print(p_acf_diff)
```




```
print(p_pacf_diff)
```




```
# Manual Model Selection (SARIMA Candidates)
# Since we observed seasonality in decomposition, we must include 'seasonal' arguments.
# We fit these to the raw data (ts_home) as decided.

# Model 1: The "Airline Model" (Classic for trend + season)
# ARIMA(0,1,1)(0,1,1)[12]
m1 <- Arima(ts_home, order=c(0,1,1), seasonal=c(0,1,1))
```

```r
# Model 2: Seasonal AR Model (Focus on autoregression)
# ARIMA(1,1,0)(1,0,0)[12]
m2 <- Arima(ts_home, order=c(1,1,0), seasonal=c(1,0,0))

# Model 3: Mixed Model
# ARIMA(1,1,1)(0,1,1)[12]
m3 <- Arima(ts_home, order=c(1,1,1), seasonal=c(0,1,1))

# AIC/BIC Comparison Table
model_comp <- data.frame(
  Model = c("ARIMA(0,1,1)(0,1,1)[12]", "ARIMA(1,1,0)(1,0,0)[12]",
    ↪  "ARIMA(1,1,1)(0,1,1)[12]"),
  AIC = c(m1$aic, m2$aic, m3$aic),
  BIC = c(m1$bic, m2$bic, m3$bic)
)

# Display table
kable(model_comp, caption = "Manual SARIMA Model Comparison") %>% kable_styling()
```

Table 1: Manual SARIMA Model Comparison

Model	AIC	BIC
ARIMA(0,1,1)(0,1,1)[12]	-3037.791	-3026.730
ARIMA(1,1,0)(1,0,0)[12]	-3227.032	-3215.851
ARIMA(1,1,1)(0,1,1)[12]	-3054.036	-3039.288

```r
# Auto ARIMA & Full Diagnostics

# 1. Run Auto ARIMA
auto_fit <- auto.arima(ts_home, seasonal = TRUE, stepwise = FALSE, approximation = FALSE)
print(summary(auto_fit))
```

```
Series: ts_home
ARIMA(3,1,2)

Coefficients:
          ar1      ar2      ar3     ma1     ma2
      -1.0784  -0.9617  -0.1857  0.7295  0.8724
s.e.   0.0840   0.0984   0.0725  0.0578  0.0576

sigma^2 = 1.49e-06:  log likelihood = 1625.94
AIC=-3239.88   AICc=-3239.6   BIC=-3217.52

Training set error measures:
                      ME        RMSE          MAE        MPE     MAPE
Training set 5.288114e-05 0.001208867 0.0006067506 0.05429668 0.893977
                  MASE        ACF1
Training set 0.1978204 0.007714963
```

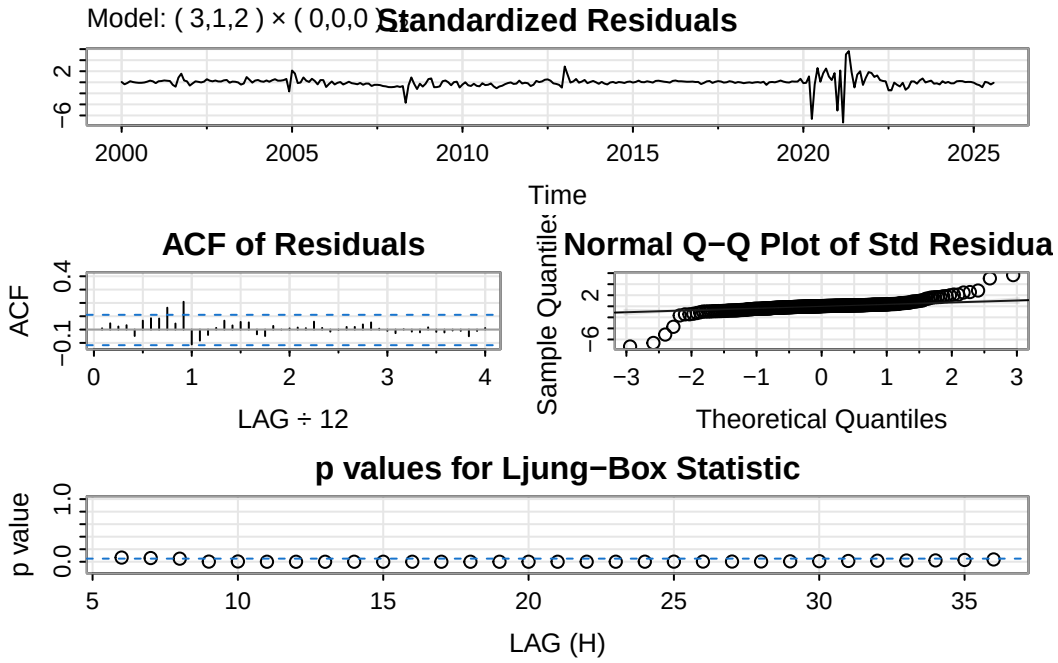
```r
# 2. Extract orders robustly
ord <- arimaorder(auto_fit)

# Check if the model has seasonal components (length > 3)
if (length(ord) > 3) {
  # Seasonal Model Found: Use extracted P, D, Q
  p <- ord[1]; d <- ord[2]; q <- ord[3]
  P <- ord[4]; D <- ord[5]; Q <- ord[6]
} else {
  # Non-seasonal Model Found: Force Seasonal parts to 0
  p <- ord[1]; d <- ord[2]; q <- ord[3]
  P <- 0;      D <- 0;      Q <- 0
}

cat("Diagnosing Model -> ",
    "ARIMA(", p, ",", d, ",", q, ")(", P, ",", D, ",", Q, ")[12]\n", sep="")
```

```
Diagnosing Model -> ARIMA(3,1,2)(0,0,0)[12]
```

```r
# 3. Full Diagnostics using sarima()
# Now passing safe, non-NA variables
sarima(ts_home, p, d, q, P, D, Q, 12)
```

```
initial  value -6.585939
iter   2 value -6.658142
iter   3 value -6.687841
iter   4 value -6.688912
iter   5 value -6.689113
iter   6 value -6.689310
iter   7 value -6.691123
iter   8 value -6.692498
iter   9 value -6.693926
iter  10 value -6.694109
iter  11 value -6.694289
iter  12 value -6.694392
iter  13 value -6.695036
iter  14 value -6.698326
iter  15 value -6.700281
iter  16 value -6.700499
iter  17 value -6.701133
iter  18 value -6.701966
iter  19 value -6.702883
iter  20 value -6.704059
iter  21 value -6.707205
iter  22 value -6.707983
iter  23 value -6.711398
iter  24 value -6.711979
iter  25 value -6.712280
iter  26 value -6.712396
iter  27 value -6.712403
iter  28 value -6.712404
iter  29 value -6.712405
```

```
iter  30 value -6.712405
iter  30 value -6.712405
iter  30 value -6.712405
final  value -6.712405
converged
initial  value -6.716079
iter   2 value -6.716094
iter   3 value -6.716113
iter   4 value -6.716122
iter   5 value -6.716129
iter   6 value -6.716129
iter   7 value -6.716130
iter   7 value -6.716130
iter   7 value -6.716130
final  value -6.716130
converged
<><><><><><><><><><><><><><><>

Coefficients:
          Estimate     SE  t.value p.value
ar1       -1.0808 0.0844 -12.8124  0.0000
ar2       -0.9645 0.0983  -9.8154  0.0000
ar3       -0.1876 0.0724  -2.5899  0.0101
ma1        0.7301 0.0581  12.5624  0.0000
ma2        0.8728 0.0575  15.1889  0.0000
constant   0.0000 0.0001   0.5377  0.5912

sigma^2 estimated as 1.463286e-06 on 301 degrees of freedom

AIC = -10.54878  AICc = -10.54787  BIC = -10.4638
```



```
### Benchmarking (ARIMA vs. Seasonal Naive)
# Revised version: ensure date formats are consistent to avoid transform_date errors

# Ensure the zoo package is available for time-series date handling
if (!require(zoo)) install.packages("zoo")
library(zoo)

# 1. Split Data (Train / Test)
h_step <- 12

train_set <- head(ts_home, length(ts_home) - h_step)
test_set  <- tail(ts_home, h_step)

# 2. Fit Models
fit_arima_train <- Arima(train_set, order = c(0, 1, 1), seasonal = c(0, 1, 1))
fc_arima <- forecast(fit_arima_train, h = h_step)

fit_snaive <- snaive(train_set, h = h_step)

# 3. Calculate Accuracy Metrics
acc_arima  <- accuracy(fc_arima, test_set)
acc_snaive <- accuracy(fit_snaive, test_set)

res_table <- rbind(
  "SARIMA (0,1,1)(0,1,1)" = acc_arima[2, c("RMSE", "MAE", "MAPE")],
  "Benchmark (SNaive)"    = acc_snaive[2, c("RMSE", "MAE", "MAPE")]
)

kable(res_table, caption = "Model Performance Comparison") %>%
  kable_styling(full_width = FALSE)
```

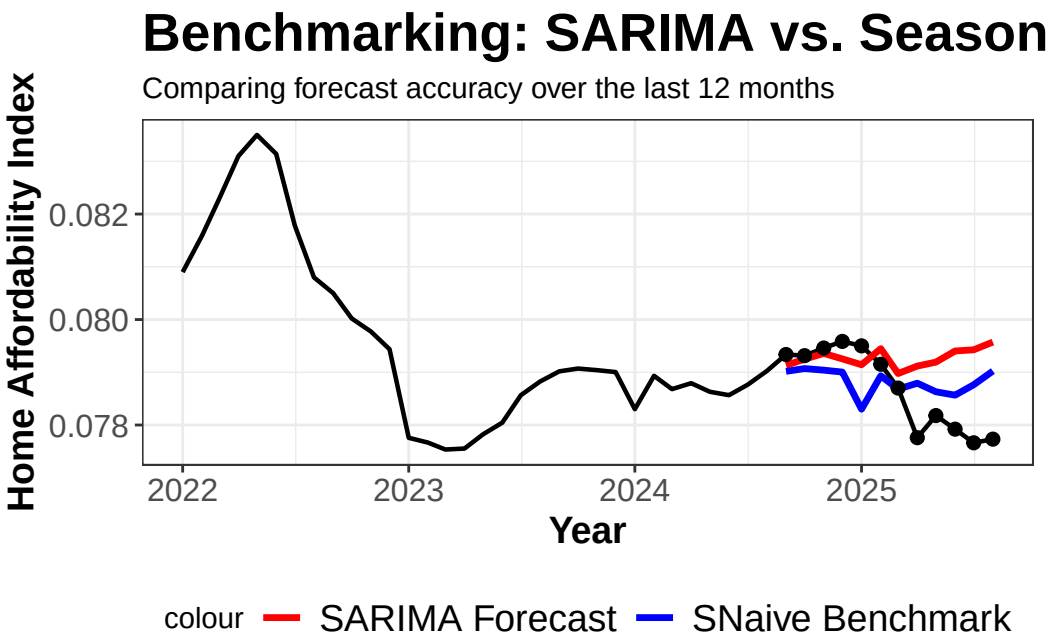
Table 2: Model Performance Comparison

	RMSE	MAE	MAPE
SARIMA (0,1,1)(0,1,1)	0.0010009	0.0007576	0.9703325
Benchmark (SNaive)	0.0007488	0.0006270	0.7992685

```
# 4. Visualization (Fixing the Date Conversion Issue)
# [Key fix]: use as.Date(time(...)) to convert decimal time indices
# into proper Date objects

df_test <- data.frame(
  Date   = as.Date(time(test_set)),
  Actual = as.numeric(test_set)
)

df_arima <- data.frame(
  Date     = as.Date(time(fc_arima$mean)),
  Forecast = as.numeric(fc_arima$mean)
)

df_snaive <- data.frame(
```

```
  Date     = as.Date(time(fit_snaive$mean)),
  Forecast = as.numeric(fit_snaive$mean)
)

# Define theme for benchmarking visualization
theme_bench <- theme_bw() +
  theme(
    plot.title = element_text(size = 20, face = "bold"),
    legend.position = "bottom",
    legend.text = element_text(size = 14),
    axis.text = element_text(size = 12),
    axis.title = element_text(size = 14, face = "bold")
  )

# Note: Instead of using autoplot(ts),
# we directly construct the plot from data frames.
# This approach is more robust and avoids date-related errors.

ggplot() +
  # Plot historical data (last 3 years only)
  geom_line(
    data = data.frame(
      Date  = as.Date(time(window(ts_home, start = end(ts_home)[1] - 3))),
      Value = as.numeric(window(ts_home, start = end(ts_home)[1] - 3))
    ),
    aes(x = Date, y = Value),
    color = "black",
    size = 0.8
  ) +

  # Plot forecast lines
  geom_line(
    data = df_arima,
    aes(x = Date, y = Forecast, color = "SARIMA Forecast"),
    size = 1.2
  ) +
  geom_line(
    data = df_snaive,
    aes(x = Date, y = Forecast, color = "SNaive Benchmark"),
    size = 1.2
  ) +

  # Plot actual test observations
  geom_point(
    data = df_test,
    aes(x = Date, y = Actual),
    color = "black",
    size = 2
  ) +

  labs(
    title = "Benchmarking: SARIMA vs. Seasonal Naive",
    subtitle = "Comparing forecast accuracy over the last 12 months",
    y = "Home Affordability Index",
```

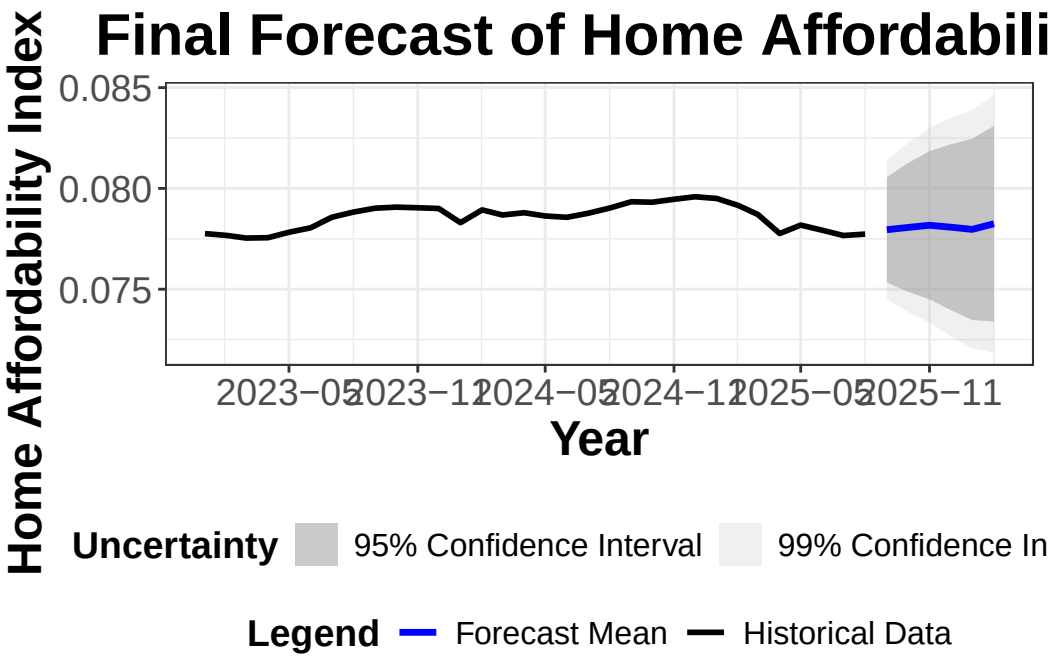
```
    x = "Year"
  ) +

  # Explicit color and date formatting
  scale_color_manual(
    values = c(
      "SARIMA Forecast"   = "red",
      "SNaive Benchmark"  = "blue"
    )
  ) +
  scale_x_date(date_labels = "%Y", date_breaks = "1 year") +

  theme_bench
```




```
### Final Forecast & Interpretation (Updated for 95% & 99% CI)

# 1. Refit the Best Model on the FULL Dataset
final_model <- Arima(ts_home, order = c(0, 1, 1), seasonal = c(0, 1, 1))

# 2. Forecast Future Values with 95% and 99% Confidence Intervals
# Key modification: specify multiple confidence levels using level = c(95, 99)
final_fc <- forecast(final_model, h = 6, level = c(95, 99))

# 3. Prepare Forecast Data for Plotting
df_final_fc <- data.frame(
  Date     = as.Date(time(final_fc$mean)),
  Forecast = as.numeric(final_fc$mean),

  # When level = c(95, 99),
  # the first column typically corresponds to 95%,
  # and the second column corresponds to 99%
```

```
  Lower95 = as.numeric(final_fc$lower[, 1]),
  Upper95 = as.numeric(final_fc$upper[, 1]),

  Lower99 = as.numeric(final_fc$lower[, 2]),
  Upper99 = as.numeric(final_fc$upper[, 2])
)

# Extract recent historical data (last 2 years) to connect with the forecast
recent_hist <- window(ts_home, start = end(ts_home)[1] - 2)

df_hist <- data.frame(
  Date  = as.Date(time(recent_hist)),
  Value = as.numeric(recent_hist)
)

ggplot() +

  # 1. Confidence Intervals
  geom_ribbon(
    data = df_final_fc,
    aes(x = Date, ymin = Lower99, ymax = Upper99, fill = "99% Confidence Interval"),
    alpha = 0.3
  ) +
  geom_ribbon(
    data = df_final_fc,
    aes(x = Date, ymin = Lower95, ymax = Upper95, fill = "95% Confidence Interval"),
    alpha = 0.5
  ) +

  # 2. Historical Data Line
  geom_line(
    data = df_hist,
    aes(x = Date, y = Value, color = "Historical Data"),
    size = 1
  ) +

  # 3. Forecast Mean Line
  geom_line(
    data = df_final_fc,
    aes(x = Date, y = Forecast, color = "Forecast Mean"),
    size = 1.2
  ) +

  # 4. Aesthetic Mappings
  scale_color_manual(
    name = "Legend",
    values = c(
      "Historical Data" = "black",
      "Forecast Mean"   = "blue"
    )
  ) +
  scale_fill_manual(
    name = "Uncertainty",
    values = c(
```

```
    "99% Confidence Interval" = "gray80",
    "95% Confidence Interval" = "gray60"
  )
) +

# 5. Labels and Theme
labs(
  title = "Final Forecast of Home Affordability",
  y = "Home Affordability Index Value",
  x = "Year"
) +
scale_x_date(date_labels = "%Y-%m", date_breaks = "6 months") +

theme_bw() +
theme(
  plot.title = element_text(size = 22, face = "bold", hjust = 0.5),
  plot.subtitle = element_text(size = 16, color = "gray40", hjust = 0.5),
  axis.title = element_text(size = 18, face = "bold"),
  axis.text = element_text(size = 14),

  # Key legend layout adjustments
  legend.position = "bottom",

  # Stack legend groups vertically
  legend.box = "vertical",

  # Center-align legend boxes
  legend.box.just = "center",

  legend.title = element_text(size = 14, face = "bold"),
  legend.text = element_text(size = 12),

  # Increase spacing between legend groups
  legend.spacing.y = unit(0.2, "cm"),
  legend.margin = margin(t = 10)
)
```



6. Time-Series

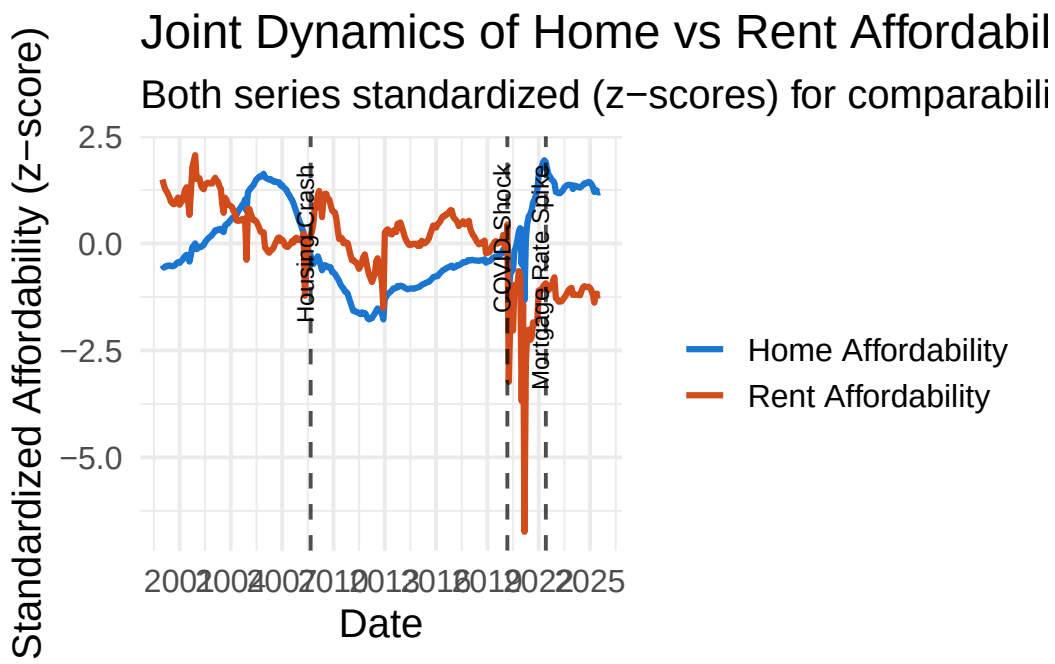
```
# Joint Dynamics of Home vs Rent Affordability
library(dplyr)
library(ggplot2)
library(scales)
library(lubridate)
library(tidyr)
library(urca)
library(broom)

# Load updated dataset
df <- read.csv("data_part/housing_affordability_monthly_final.csv")
```

```
# Convert date column (already YYYY-MM-DD format)
df$date <- as.Date(df$date)
```

```
# Standardize affordability indexes (z-scores)
df <- df %>%
  mutate(
    home_aff_z = as.numeric(scale(home_aff)),
    rent_aff_z = as.numeric(scale(rent_aff))
  )
```

```
# Joint Dynamics Plot
joint_dynamics <- ggplot(df, aes(x = date)) +
  geom_line(aes(y = home_aff_z, color = "Home Affordability"), linewidth = 1.1) +
  geom_line(aes(y = rent_aff_z, color = "Rent Affordability"), linewidth = 1.1) +
```

```r
  geom_vline(xintercept = as.Date("2008-09-01"),
             linetype = "dashed", color = "black", alpha = 0.7) +
  geom_vline(xintercept = as.Date("2020-03-01"),
             linetype = "dashed", color = "black", alpha = 0.7) +
  geom_vline(xintercept = as.Date("2022-06-01"),
             linetype = "dashed", color = "black", alpha = 0.7) +

  annotate("text", x = as.Date("2008-09-01") - 120,
           y = max(df$home_aff_z, na.rm = TRUE) * 0.95,
           label = "Housing Crash", angle = 90, size = 3, hjust = 1) +
  annotate("text", x = as.Date("2020-03-01") - 120,
           y = max(df$home_aff_z, na.rm = TRUE) * 0.95,
           label = "COVID Shock", angle = 90, size = 3, hjust = 1) +
  annotate("text", x = as.Date("2022-06-01") - 120,
           y = max(df$home_aff_z, na.rm = TRUE) * 0.95,
           label = "Mortgage Rate Spike", angle = 90, size = 3, hjust = 1) +

  labs(
    title = "Joint Dynamics of Home vs Rent Affordability",
    subtitle = "Both series standardized (z-scores) for comparability",
    x = "Date",
    y = "Standardized Affordability (z-score)",
    color = ""
  ) +
  scale_color_manual(values = c(
    "Home Affordability" = "#1B76D1",
    "Rent Affordability" = "#D14B1B"
  )) +
  scale_x_date(date_labels = "%Y", date_breaks = "3 years") +
  theme_minimal(base_size = 15)

print(joint_dynamics)
```




```
# Prepare differenced series (dh, dr)
h <- na.omit(df$home_aff)
r <- na.omit(df$rent_aff)

dh <- diff(h)
dr <- diff(r)

# ADF Tests
adf_home <- ur.df(df$home_aff, type = "drift", selectlags = "AIC")
adf_rent <- ur.df(df$rent_aff, type = "drift", selectlags = "AIC")

cat("\nADF Test for HomeAff (level):\n")
```

```
ADF Test for HomeAff (level):
```

```
summary(adf_home)
```

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
```

```
       Min         1Q     Median         3Q        Max
-0.0102207 -0.0003418  0.0000159  0.0003152  0.0072364

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0005342  0.0006472   0.825    0.410
z.lag.1     -0.0069036  0.0093985  -0.735    0.463
z.diff.lag  -0.4012814  0.0527257  -7.611 3.48e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001262 on 303 degrees of freedom
Multiple R-squared:  0.1657,    Adjusted R-squared:  0.1602
F-statistic: 30.09 on 2 and 303 DF,  p-value: 1.202e-12


Value of test-statistic is: -0.7345 0.6361

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

```
cat("\nADF Test for RentAff (level):\n")
```

```
ADF Test for RentAff (level):
```

```
summary(adf_rent)
```

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0119261 -0.0001887  0.0000182  0.0002793  0.0066348

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.006263   0.002217   2.826  0.00503 **
z.lag.1     -0.077454   0.027239  -2.843  0.00477 **
z.diff.lag  -0.480356   0.050313  -9.547  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.00126 on 303 degrees of freedom
Multiple R-squared:  0.2881,    Adjusted R-squared:  0.2834
F-statistic: 61.32 on 2 and 303 DF,  p-value: < 2.2e-16


Value of test-statistic is: -2.8435 4.1669

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

```r
# Differenced ADF
adf_dh <- ur.df(dh, type = "drift", selectlags = "AIC")
adf_dr <- ur.df(dr, type = "drift", selectlags = "AIC")

cat("\nADF Test for Differenced HomeAff:\n")
```

```
ADF Test for Differenced HomeAff:
```

```r
summary(adf_dh)
```

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0097188 -0.0003027  0.0000188  0.0002836  0.0082683

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.490e-05  7.205e-05   0.762   0.4467
z.lag.1     -1.236e+00  9.576e-02 -12.907   <2e-16 ***
z.diff.lag  -1.205e-01  5.711e-02  -2.110   0.0356 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001256 on 302 degrees of freedom
Multiple R-squared:  0.707, Adjusted R-squared:  0.705
F-statistic: 364.3 on 2 and 302 DF,  p-value: < 2.2e-16
```

```
Value of test-statistic is: -12.9067 83.2913

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

```
cat("\nADF Test for Differenced RentAff:\n")
```

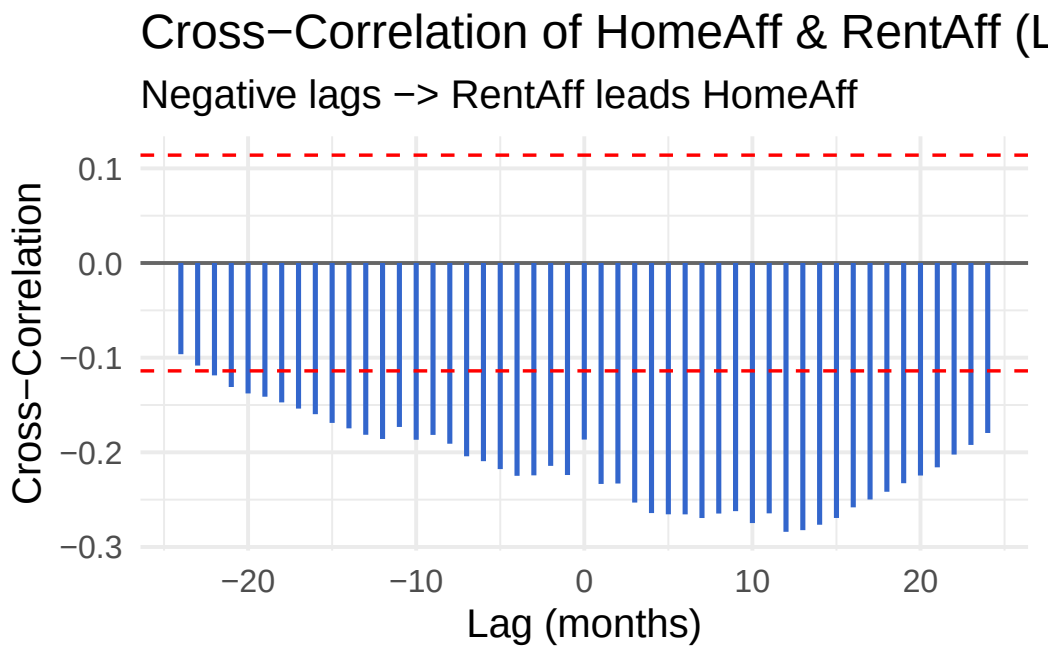
```
ADF Test for Differenced RentAff:
```

```
summary(adf_dr)
```

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift


Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0117485 -0.0002617  0.0000145  0.0002831  0.0069655

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.729e-05  7.303e-05  -0.511    0.610
z.lag.1     -1.635e+00  9.998e-02 -16.355   <2e-16 ***
z.diff.lag   7.644e-02  5.736e-02   1.333    0.184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001275 on 302 degrees of freedom
Multiple R-squared:  0.7609,    Adjusted R-squared:  0.7593
F-statistic: 480.6 on 2 and 302 DF,  p-value: < 2.2e-16


Value of test-statistic is: -16.3547 133.7377

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

```
# Use z-scored affordability (trend-preserving)
h <- df$home_aff_z
r <- df$rent_aff_z
```

```r
# Remove NA in case early periods differ
valid <- complete.cases(h, r)
h <- h[valid]
r <- r[valid]

# Compute CCF on level series
cc <- ccf(h, r, plot = FALSE, lag.max = 24)

cc_df <- data.frame(
  lag = cc$lag,
  ccf = cc$acf
)

# Significance threshold (same formula)
crit <- 2 / sqrt(length(h))

# Plot
p2 <- ggplot(cc_df, aes(x = lag, y = ccf)) +
  geom_hline(yintercept = 0, color = "gray40") +
  geom_segment(
    aes(x = lag, xend = lag, y = 0, yend = ccf),
    linewidth = 0.9, color = "#3366CC"
  ) +
  geom_hline(yintercept = crit,
             linetype = "dashed", color = "red", linewidth = 0.6) +
  geom_hline(yintercept = -crit,
             linetype = "dashed", color = "red", linewidth = 0.6) +
  labs(
    title = "Cross-Correlation of HomeAff & RentAff (Level Series)",
    subtitle = "Negative lags → RentAff leads HomeAff",
    x = "Lag (months)",
    y = "Cross-Correlation"
  ) +
  theme_minimal(base_size = 15)

print(p2)
```




```r
# Cross-lag regression (Rent leads Home)
lag_df <- data.frame(
  dh = dh,
  dr_lag1 = dplyr::lag(dr, 1),
  dr_lag2 = dplyr::lag(dr, 2),
  dr_lag3 = dplyr::lag(dr, 3),
  dr_lag6 = dplyr::lag(dr, 6),
  dr_lag9 = dplyr::lag(dr, 9),
  dr_lag12 = dplyr::lag(dr, 12)
)

lag_df <- na.omit(lag_df)

model_rent_leads <- lm(dh ~ dr_lag1 + dr_lag2 + dr_lag3 +
                         dr_lag6 + dr_lag9 + dr_lag12,
                       data = lag_df)

summary(model_rent_leads)
```

```
Call:
lm(formula = dh ~ dr_lag1 + dr_lag2 + dr_lag3 + dr_lag6 + dr_lag9 +
    dr_lag12, data = lag_df)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0100518 -0.0003153  0.0000448  0.0002825  0.0070304

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.536e-05  7.022e-05   0.504  0.61499
dr_lag1     -4.616e-01  5.623e-02  -8.209 7.53e-15 ***
```

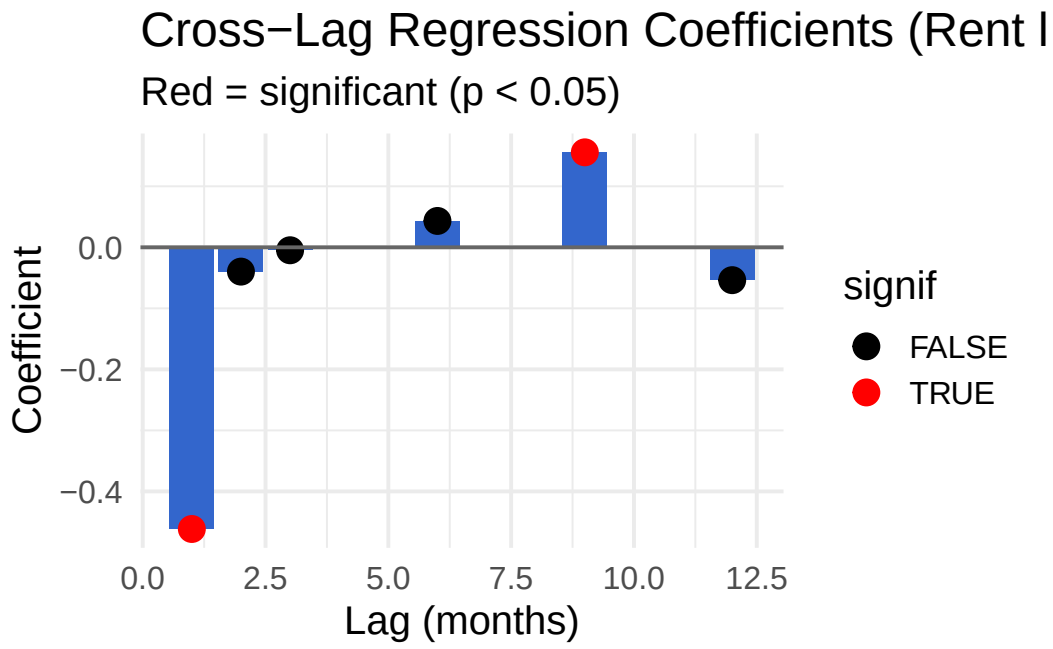
```
dr_lag2     -3.981e-02  6.339e-02  -0.628  0.53053
dr_lag3     -4.817e-03  5.526e-02  -0.087  0.93059
dr_lag6      4.324e-02  4.744e-02   0.912  0.36279
dr_lag9      1.557e-01  4.702e-02   3.310  0.00105 **
dr_lag12    -5.380e-02  4.959e-02  -1.085  0.27880
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001205 on 288 degrees of freedom
Multiple R-squared:  0.2774,    Adjusted R-squared:  0.2624
F-statistic: 18.43 on 6 and 288 DF,  p-value: < 2.2e-16
```

```r
# Cross-lag regression (Home leads Rent)
lag_df2 <- data.frame(
  dr = dr,
  dh_lag1 = dplyr::lag(dh, 1),
  dh_lag2 = dplyr::lag(dh, 2),
  dh_lag3 = dplyr::lag(dh, 3),
  dh_lag6 = dplyr::lag(dh, 6),
  dh_lag9 = dplyr::lag(dh, 9),
  dh_lag12 = dplyr::lag(dh, 12)
)

lag_df2 <- na.omit(lag_df2)

model_home_leads <- lm(dr ~ dh_lag1 + dh_lag2 + dh_lag3 +
                         dh_lag6 + dh_lag9 + dh_lag12,
                       data = lag_df2)

summary(model_home_leads)
```

```
Call:
lm(formula = dr ~ dh_lag1 + dh_lag2 + dh_lag3 + dh_lag6 + dh_lag9 +
    dh_lag12, data = lag_df2)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0112926 -0.0002370  0.0000165  0.0002442  0.0072873

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.207e-05  7.565e-05  -0.160 0.873381
dh_lag1     -5.417e-01  6.134e-02  -8.832  < 2e-16 ***
dh_lag2      2.747e-02  6.265e-02   0.438 0.661372
dh_lag3      4.292e-02  5.990e-02   0.716 0.474272
dh_lag6      6.823e-02  5.430e-02   1.257 0.209897
dh_lag9      2.040e-01  5.436e-02   3.753 0.000212 ***
dh_lag12    -3.326e-02  5.714e-02  -0.582 0.560909
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001294 on 288 degrees of freedom
Multiple R-squared:  0.2861,    Adjusted R-squared:  0.2712
```

```
F-statistic: 19.24 on 6 and 288 DF,  p-value: < 2.2e-16
```

```r
# Coefficient plot (Rent leads Home)
coef_df <- broom::tidy(model_rent_leads) %>%
  filter(term != "(Intercept)") %>%
  mutate(
    lag = as.numeric(gsub("dr_lag", "", term)),
    signif = p.value < 0.05
  )

ggplot(coef_df, aes(x = lag, y = estimate)) +
  geom_col(fill = "#3366CC") +
  geom_point(aes(color = signif), size = 4) +
  scale_color_manual(values = c("black", "red")) +
  geom_hline(yintercept = 0, color = "gray40") +
  labs(
    title = "Cross-Lag Regression Coefficients (Rent leads Home)",
    subtitle = "Red = significant (p < 0.05)",
    x = "Lag (months)",
    y = "Coefficient"
  ) +
  theme_minimal(base_size = 15)
```



7.Regression Results

```r
## Step 0: Load packages

library(tseries)
library(lmtest)
```

```r
library(knitr)
library(ggplot2)
library(broom)

## Step 1: Read and prepare data

# Read final dataset
housing <- read.csv("data_part/housing_affordability_monthly_final.csv",
                    stringsAsFactors = FALSE)

# Original date is "YYYY-MM" → add "-01" to create a proper Date
housing$date <- as.Date(paste0(housing$date, "-01"))

# Make sure data are sorted by time
housing <- housing[order(housing$date), ]

# Quick check of key variables
str(housing[, c("home_aff", "mortgage_rate", "wage_income")])
```

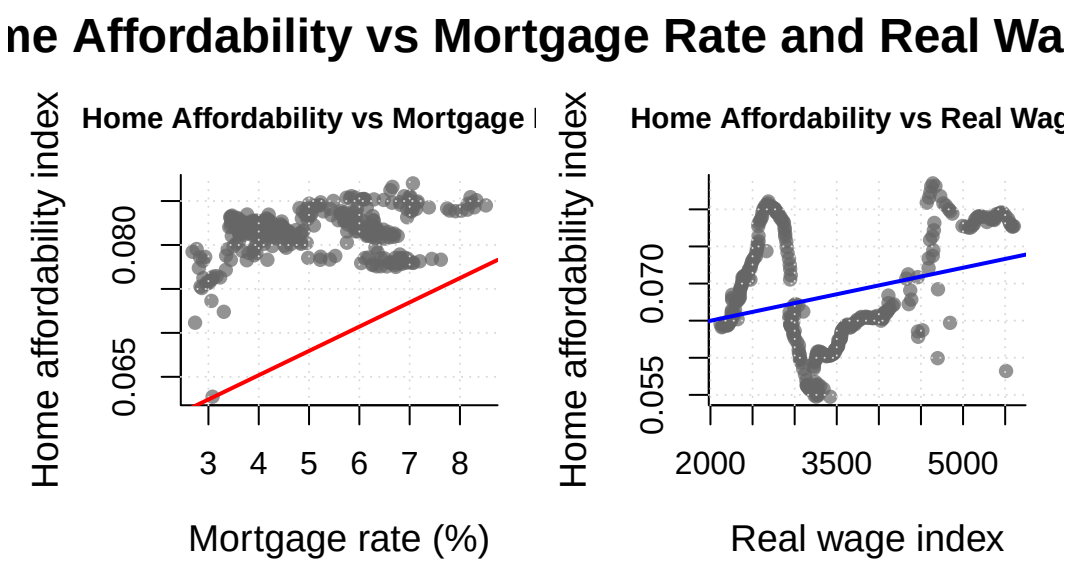
```
'data.frame':   308 obs. of  3 variables:
 $ home_aff     : num  0.0646 0.0641 0.0642 0.0644 0.0644 ...
 $ mortgage_rate: num  8.21 8.32 8.24 8.15 8.52 ...
 $ wage_income  : num  2120 2133 2146 2153 2162 ...
```

```r
summary(housing[, c("home_aff", "mortgage_rate", "wage_income")])
```

```
    home_aff       mortgage_rate    wage_income
 Min.   :0.05475   Min.   :2.684   Min.   :2120
 1st Qu.:0.06312   1st Qu.:3.957   1st Qu.:2771
 Median :0.06615   Median :5.053   Median :3258
 Mean   :0.06848   Mean   :5.192   Mean   :3464
 3rd Qu.:0.07649   3rd Qu.:6.333   3rd Qu.:4052
 Max.   :0.08350   Max.   :8.515   Max.   :5603
```

```r
## Step 2: Check if linear regression is roughly appropriate
## Goal: See whether home_aff roughly has linear relationships
## with mortgage_rate and wage_income.

# Pairwise scatterplots
par(mfrow = c(1, 2),
    mar = c(4.5, 4.5, 3, 1),
    oma = c(0, 0, 4, 0))

pt_col <- rgb(0.4, 0.4, 0.4, alpha = 0.7)

# (a) vs mortgage rate
plot(housing$mortgage_rate, housing$rent_aff,
     pch = 16, cex = 1, col = pt_col,
     xlab = "Mortgage rate (%)",
     ylab = "Home affordability index",
     main = "Home Affordability vs Mortgage Rate",
     cex.main = 0.9,
```

```
     cex.lab = 1.2,
     cex.axis = 1,
     bty = "l")
grid(col = "grey85")
abline(lm(home_aff ~ mortgage_rate, data = housing),
       col = "red", lwd = 2)

# (b) vs real wages
plot(housing$wage_income, housing$home_aff,
     pch = 16, cex = 1, col = pt_col,
     xlab = "Real wage index",
     ylab = "Home affordability index",
     main = "Home Affordability vs Real Wages",
     cex.main = 0.9,
     cex.lab = 1.2,
     cex.axis = 1,
     bty = "l")
grid(col = "grey85")
abline(lm(home_aff ~ wage_income, data = housing),
       col = "blue", lwd = 2)

# Overall title
mtext("Home Affordability vs Mortgage Rate and Real Wages",
      outer = TRUE, cex = 1.5, font = 2)
```




```
par(mfrow = c(1, 1), oma = c(0,0,0,0))

# Pearson correlation matrix
cor(housing[, c("home_aff", "mortgage_rate", "wage_income")])
```

```
              home_aff mortgage_rate wage_income
```

```
home_aff      1.0000000     0.5001435   0.2896075
mortgage_rate 0.5001435     1.0000000  -0.2424641
wage_income   0.2896075    -0.2424641   1.0000000
```

```
## Step 3: Time-series properties - ADF tests (stationarity)
## ADF null hypothesis: series has a unit root (non-stationary).
## If p-value < 0.05 → reject H0 → series is stationary.
## If p-value >= 0.05 → fail to reject H0 → treat as non-stationary.

adf_home <- adf.test(housing$home_aff, k = 12)
adf_rate <- adf.test(housing$mortgage_rate, k = 12)
adf_wage <- adf.test(housing$wage_income, k = 12)

adf_home
```

```
    Augmented Dickey-Fuller Test

data:  housing$home_aff
Dickey-Fuller = -2.1642, Lag order = 12, p-value = 0.5071
alternative hypothesis: stationary
```

```
adf_rate
```

```
    Augmented Dickey-Fuller Test

data:  housing$mortgage_rate
Dickey-Fuller = -1.5469, Lag order = 12, p-value = 0.7674
alternative hypothesis: stationary
```

```
adf_wage
```

```
    Augmented Dickey-Fuller Test

data:  housing$wage_income
Dickey-Fuller = -0.077106, Lag order = 12, p-value = 0.99
alternative hypothesis: stationary
```

```
## Step 4: Fit the main OLS regression
## Dependent variable: home_aff (housing affordability index)
## Key regressors: mortgage_rate, wage_income
## This is a simple association, not a causal structural model.

mod_home <- lm(home_aff ~ mortgage_rate + wage_income,
               data = housing)

summary(mod_home)
```

```
Call:
lm(formula = home_aff ~ mortgage_rate + wage_income, data = housing)

Residuals:
      Min        1Q    Median        3Q       Max
-0.010537 -0.003733 -0.001488  0.004054  0.014310

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.867e-02  2.035e-03  19.000   <2e-16 ***
mortgage_rate 3.347e-03  2.462e-04  13.594   <2e-16 ***
wage_income   3.587e-06  3.663e-07   9.792   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.005854 on 305 degrees of freedom
Multiple R-squared:  0.4295,    Adjusted R-squared:  0.4258
F-statistic: 114.8 on 2 and 305 DF,  p-value: < 2.2e-16
```

```r
## Step 5: Build a clean regression table

reg_sum  <- summary(mod_home)
coef_tab <- coef(reg_sum)  # matrix with Estimate, Std. Error, t value, Pr(>|t|)

# Extract R-squared and adjusted R-squared
r2     <- reg_sum$r.squared
adj_r2 <- reg_sum$adj.r.squared

kable(round(coef_tab, 4),
      caption = "Regression of home affordability on mortgage rate and wages")
```

Table 3: Regression of home affordability on mortgage rate and wages

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0387	2e-03	19.0000	0
mortgage_rate	0.0033	2e-04	13.5935	0
wage_income	0.0000	0e+00	9.7924	0

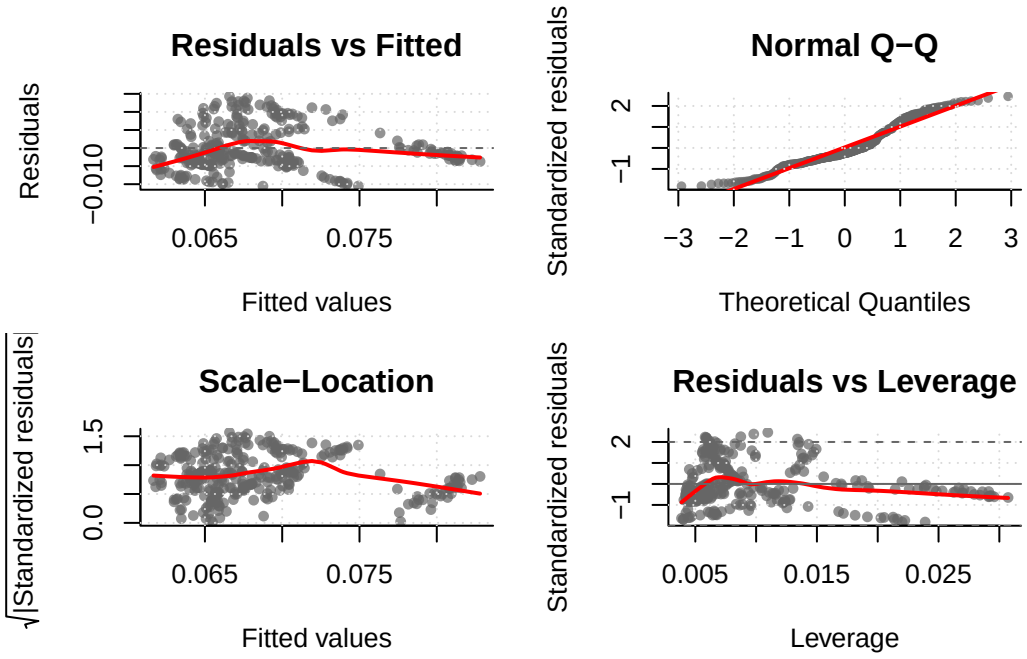
```r
cat(sprintf("R-squared: %.3f\nAdjusted R-squared: %.3f\n", r2, adj_r2))
```

```
R-squared: 0.430
Adjusted R-squared: 0.426
```

```r
## Step 6: Diagnostics for the regression model
##  (1) Residual plots → linearity, equal variance, outliers
##  (2) Residuals over time → any obvious patterns?
##  (3) Autocorrelation in residuals → ACF + Durbin-Watson
##  (4) Normality of residuals → QQ plot + Shapiro-Wilk test

# 6.1 Basic residual plots (linearity, equal variance, influential points)
```

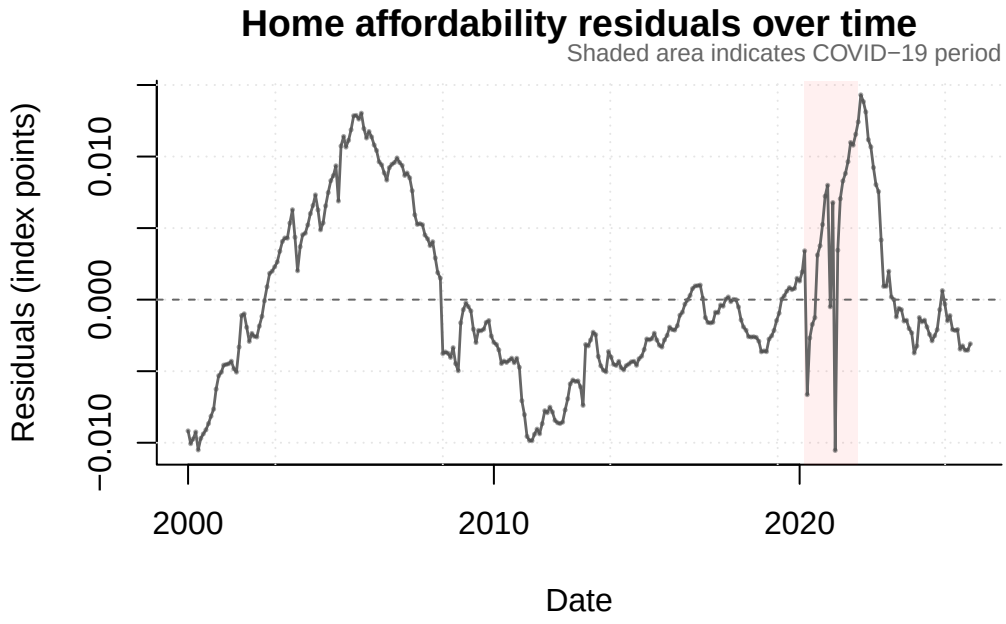
```r
# Common aesthetics (keep identical to rent)
pt_col   <- rgb(0.4, 0.4, 0.4, 0.7)
line_col <- "red"

# Extract quantities
fitted_home <- fitted(mod_home)
resid_home  <- resid(mod_home)
std_resid   <- rstandard(mod_home)
lev         <- hatvalues(mod_home)

par(mfrow = c(2, 2),
    mar = c(4.5, 4.5, 3, 1))

## (1) Residuals vs Fitted
plot(fitted_home, resid_home,
     pch = 16, cex = 0.9, col = pt_col,
     xlab = "Fitted values",
     ylab = "Residuals",
     main = "Residuals vs Fitted",
     bty = "l")
grid(col = "grey85")
lines(lowess(fitted_home, resid_home),
      col = line_col, lwd = 2)
abline(h = 0, lty = 2, col = "grey40")

## (2) Normal Q-Q
qqnorm(std_resid,
       pch = 16, cex = 0.8, col = pt_col,
       main = "Normal Q-Q",
       xlab = "Theoretical Quantiles",
       ylab = "Standardized residuals",
       bty = "l")
qqline(std_resid, col = line_col, lwd = 2)
grid(col = "grey85")

## (3) Scale-Location
plot(fitted_home, sqrt(abs(std_resid)),
     pch = 16, cex = 0.9, col = pt_col,
     xlab = "Fitted values",
     ylab = expression(sqrt("|Standardized residuals|")),
     main = "Scale-Location",
     bty = "l")
grid(col = "grey85")
lines(lowess(fitted_home, sqrt(abs(std_resid))),
      col = line_col, lwd = 2)

## (4) Residuals vs Leverage
plot(lev, std_resid,
     pch = 16, cex = 0.9, col = pt_col,
     xlab = "Leverage",
     ylab = "Standardized residuals",
     main = "Residuals vs Leverage",
     bty = "l")
grid(col = "grey85")
```

```
lines(lowess(lev, std_resid),
      col = line_col, lwd = 2)
abline(h = c(-2, 0, 2),
       lty = c(2, 1, 2),
       col = "grey40")
```

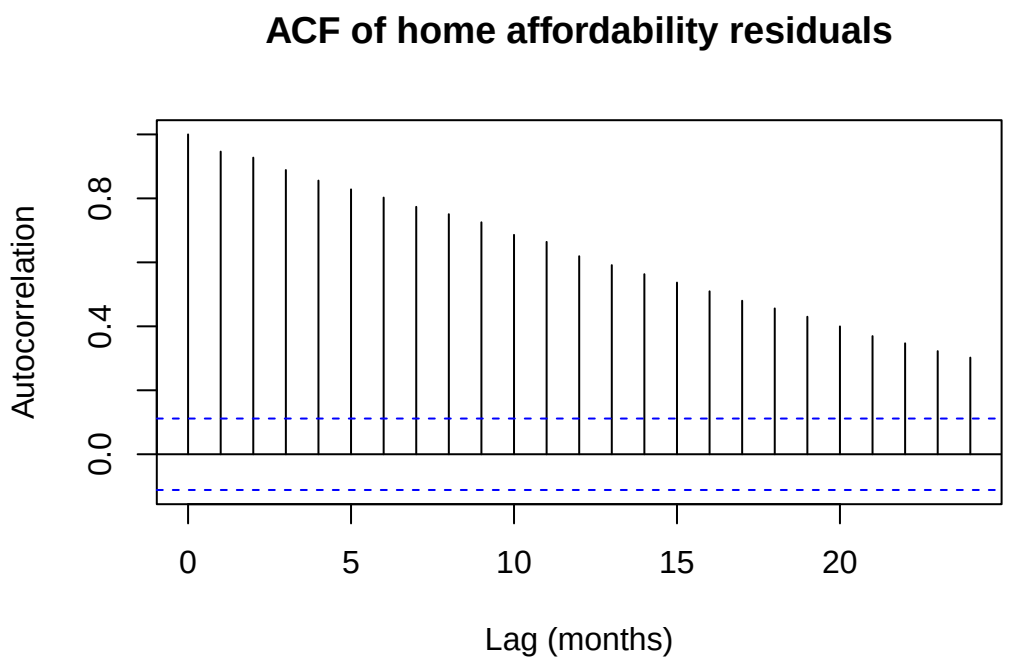



```
par(mfrow = c(1, 1))

res_home <- resid(mod_home)

covid_start <- as.Date("2020-03-01")
covid_end   <- as.Date("2021-12-01")

par(mar = c(4.5, 4.5, 3, 1))

plot(housing$date, res_home,
     type = "n",
     xlab = "Date",
     ylab = "Residuals (index points)",
     main = "Home affordability residuals over time",
     bty = "l")

grid(col = "grey90")

# Highlight COVID period
usr <- par("usr")
rect(covid_start, usr[3], covid_end, usr[4],
     col = rgb(1, 0, 0, 0.06), border = NA)

# Emphasize structure first, noise second
lines(housing$date, res_home,
```

```
      col = "grey40", lwd = 1.4)

points(housing$date, res_home,
       pch = 16, cex = 0.35,
       col = rgb(0.3, 0.3, 0.3, 0.5))

abline(h = 0, lty = 2, col = "grey40")

mtext("Shaded area indicates COVID-19 period",
      side = 3, line = 0.3, adj = 1,
      cex = 0.75, col = "grey40")
```



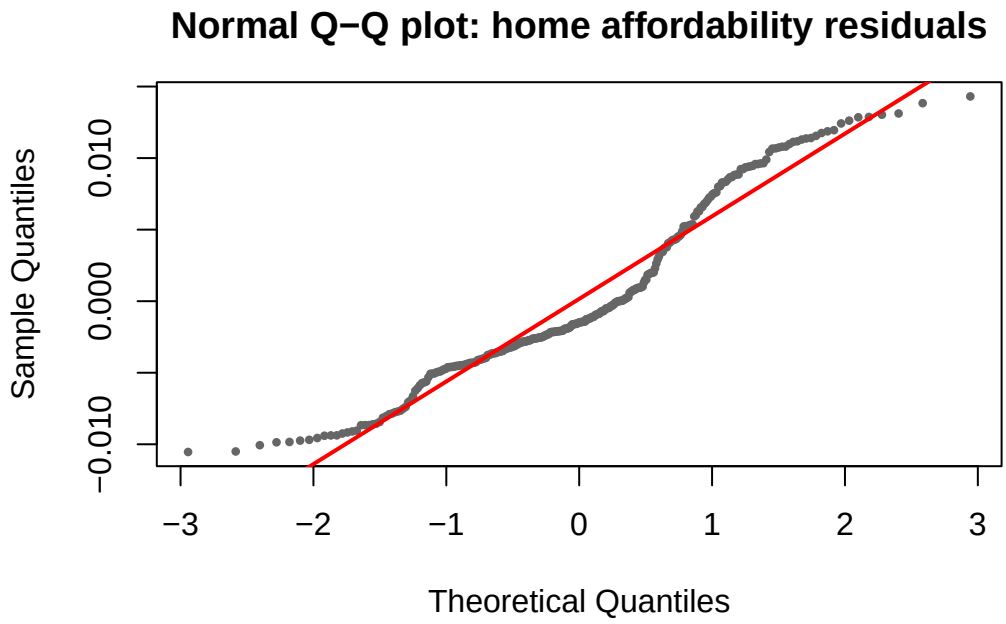

```
# 6.3 Autocorrelation diagnostics
acf(resid(mod_home),
    main = "ACF of home affordability residuals",
    xlab = "Lag (months)",
    ylab = "Autocorrelation")
```




```
# Durbin-Watson test for first-order autocorrelation
# H0: no first-order autocorrelation in errors
dwtest(mod_home)
```

```
    Durbin-Watson test

data:  mod_home
DW = 0.09856, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

```
# 6.4 Normality diagnostics
qqnorm(resid(mod_home),
       main = "Normal Q-Q plot: home affordability residuals",
       pch = 16, cex = 0.6, col = "grey40")
qqline(resid(mod_home), col = "red", lwd = 2)
```

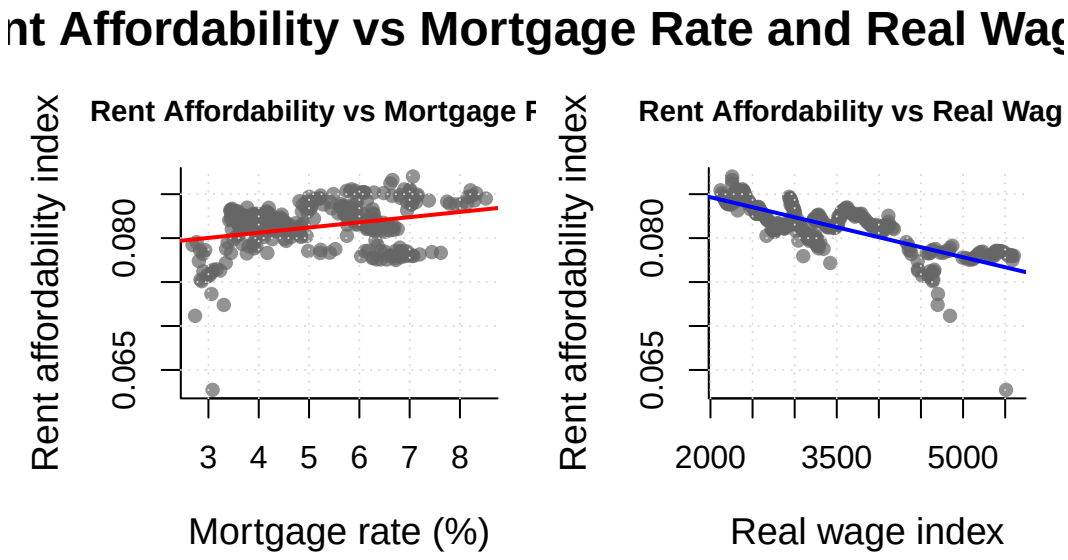



```
# Shapiro-Wilk test of normality (large n → very sensitive)
# H0: residuals are normally distributed
shapiro.test(resid(mod_home))
```

```
    Shapiro-Wilk normality test

data:  resid(mod_home)
W = 0.94497, p-value = 2.599e-09
```

```
## Step 2 (rent): Check if linear regression is roughly appropriate
## Goal: See whether rent_aff roughly has linear relationships
## with mortgage_rate and wage_income.

# Pairwise scatterplots
par(mfrow = c(1, 2),
    mar = c(4.5, 4.5, 3, 1),
    oma = c(0, 0, 4, 0))

pt_col <- rgb(0.4, 0.4, 0.4, alpha = 0.7)

# (a) vs mortgage rate
plot(housing$mortgage_rate, housing$rent_aff,
     pch = 16, cex = 1, col = pt_col,
     xlab = "Mortgage rate (%)",
     ylab = "Rent affordability index",
     main = "Rent Affordability vs Mortgage Rate",
     cex.main = 0.9,
     cex.lab = 1.2,
     cex.axis = 1,
     bty = "l")
```

```r
grid(col = "grey85")
abline(lm(rent_aff ~ mortgage_rate, data = housing),
       col = "red", lwd = 2)

# (b) vs real wages
plot(housing$wage_income, housing$rent_aff,
     pch = 16, cex = 1, col = pt_col,
     xlab = "Real wage index",
     ylab = "Rent affordability index",
     main = "Rent Affordability vs Real Wages",
     cex.main = 0.9,
     cex.lab = 1.2,
     cex.axis = 1,
     bty = "l")
grid(col = "grey85")
abline(lm(rent_aff ~ wage_income, data = housing),
       col = "blue", lwd = 2)

# Overall title
mtext("Rent Affordability vs Mortgage Rate and Real Wages",
      outer = TRUE, cex = 1.5, font = 2)
```




```r
par(mfrow = c(1, 1), oma = c(0,0,0,0))


# Pearson correlation matrix
cor(housing[, c("rent_aff", "mortgage_rate", "wage_income")])
```

```
              rent_aff mortgage_rate wage_income
rent_aff     1.0000000     0.2997523  -0.7762402
```

```
mortgage_rate  0.2997523     1.0000000  -0.2424641
wage_income   -0.7762402    -0.2424641   1.0000000
```

```r
## Step 3 (rent): Time-series properties - ADF tests (stationarity)
## ADF null hypothesis: series has a unit root (non-stationary).
## If p-value < 0.05 → reject H0 → series is stationary.
## If p-value >= 0.05 → fail to reject H0 → treat as non-stationary.

adf_rent <- adf.test(housing$rent_aff, k = 12)
adf_rate <- adf.test(housing$mortgage_rate, k = 12)
adf_wage <- adf.test(housing$wage_income, k = 12)

adf_rent
```

```
    Augmented Dickey-Fuller Test

data:  housing$rent_aff
Dickey-Fuller = -3.5463, Lag order = 12, p-value = 0.0386
alternative hypothesis: stationary
```

```r
adf_rate
```

```
    Augmented Dickey-Fuller Test

data:  housing$mortgage_rate
Dickey-Fuller = -1.5469, Lag order = 12, p-value = 0.7674
alternative hypothesis: stationary
```

```r
adf_wage
```

```
    Augmented Dickey-Fuller Test

data:  housing$wage_income
Dickey-Fuller = -0.077106, Lag order = 12, p-value = 0.99
alternative hypothesis: stationary
```

```r
## Step 4 (rent): Fit the main OLS regression
## Dependent variable: rent_aff (rent affordability index)
## Key regressors: mortgage_rate, wage_income
## This is a simple association, not a causal structural model.

mod_rent <- lm(rent_aff ~ mortgage_rate + wage_income,
               data = housing)

summary(mod_rent)
```

```
Call:
lm(formula = rent_aff ~ mortgage_rate + wage_income, data = housing)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0136048 -0.0009151  0.0001546  0.0011847  0.0030546

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    8.770e-02  5.961e-04 147.113  < 2e-16 ***
mortgage_rate  2.336e-04  7.212e-05   3.239  0.00133 **
wage_income   -2.192e-06  1.073e-07 -20.432  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001715 on 305 degrees of freedom
Multiple R-squared:  0.6158,    Adjusted R-squared:  0.6132
F-statistic: 244.4 on 2 and 305 DF,  p-value: < 2.2e-16
```

```
## Step 5 (rent): Build a clean regression table

reg_rent_sum  <- summary(mod_rent)
coef_rent_tab <- coef(reg_rent_sum)  # matrix with Estimate, Std. Error, t value,
 ↪  Pr(>|t|)

# Extract R-squared and adjusted R-squared
r2_rent     <- reg_rent_sum$r.squared
adj_r2_rent <- reg_rent_sum$adj.r.squared

kable(round(coef_rent_tab, 4),
      caption = "Regression of rent affordability on mortgage rate and wages")
```

Table 4: Regression of rent affordability on mortgage rate and wages

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0877	6e-04	147.1131	0.0000
mortgage_rate	0.0002	1e-04	3.2393	0.0013
wage_income	0.0000	0e+00	-20.4320	0.0000

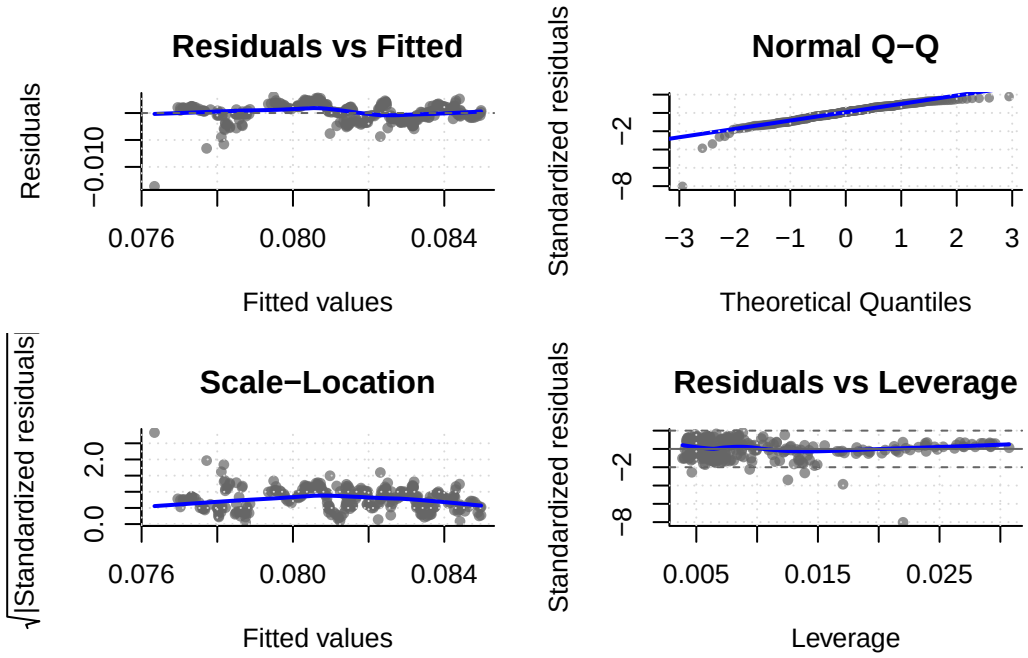
```
cat(sprintf("R-squared: %.3f\nAdjusted R-squared: %.3f\n",
            r2_rent, adj_r2_rent))
```

```
R-squared: 0.616
Adjusted R-squared: 0.613
```

```
## Step 6 (rent): Diagnostics for the regression model
##  (1) Residual plots → linearity, equal variance, outliers
##  (2) Residuals over time → any obvious patterns?
##  (3) Autocorrelation in residuals → ACF + Durbin-Watson
##  (4) Normality of residuals → QQ plot + Shapiro-Wilk test
```

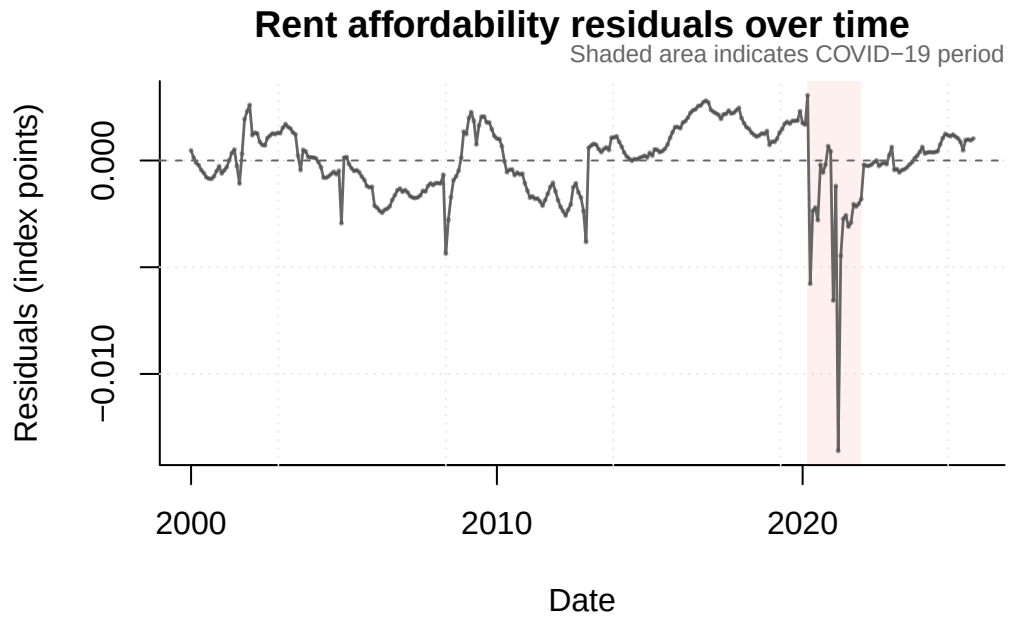
```r
# 6.1 Basic residual plots (linearity, equal variance, influential points)
# Common aesthetics
pt_col  <- rgb(0.4, 0.4, 0.4, 0.7)
line_col <- "blue"

# Extract quantities
fitted_rent <- fitted(mod_rent)
resid_rent  <- resid(mod_rent)
std_resid   <- rstandard(mod_rent)
lev         <- hatvalues(mod_rent)

par(mfrow = c(2, 2),
    mar = c(4.5, 4.5, 3, 1))

## (1) Residuals vs Fitted
plot(fitted_rent, resid_rent,
     pch = 16, cex = 0.9, col = pt_col,
     xlab = "Fitted values",
     ylab = "Residuals",
     main = "Residuals vs Fitted",
     bty = "l")
grid(col = "grey85")
lines(lowess(fitted_rent, resid_rent),
      col = line_col, lwd = 2)
abline(h = 0, lty = 2, col = "grey40")

## (2) Normal Q-Q
qqnorm(std_resid,
       pch = 16, cex = 0.8, col = pt_col,
       main = "Normal Q-Q",
       xlab = "Theoretical Quantiles",
       ylab = "Standardized residuals",
       bty = "l")
qqline(std_resid, col = line_col, lwd = 2)
grid(col = "grey85")

## (3) Scale-Location
plot(fitted_rent, sqrt(abs(std_resid)),
     pch = 16, cex = 0.9, col = pt_col,
     xlab = "Fitted values",
     ylab = expression(sqrt("|Standardized residuals|")),
     main = "Scale-Location",
     bty = "l")
grid(col = "grey85")
lines(lowess(fitted_rent, sqrt(abs(std_resid))),
      col = line_col, lwd = 2)

## (4) Residuals vs Leverage
plot(lev, std_resid,
     pch = 16, cex = 0.9, col = pt_col,
     xlab = "Leverage",
     ylab = "Standardized residuals",
     main = "Residuals vs Leverage",
     bty = "l")
```

```
grid(col = "grey85")
lines(lowess(lev, std_resid),
      col = line_col, lwd = 2)
abline(h = c(-2, 0, 2),
       lty = c(2, 1, 2),
       col = "grey40")
```

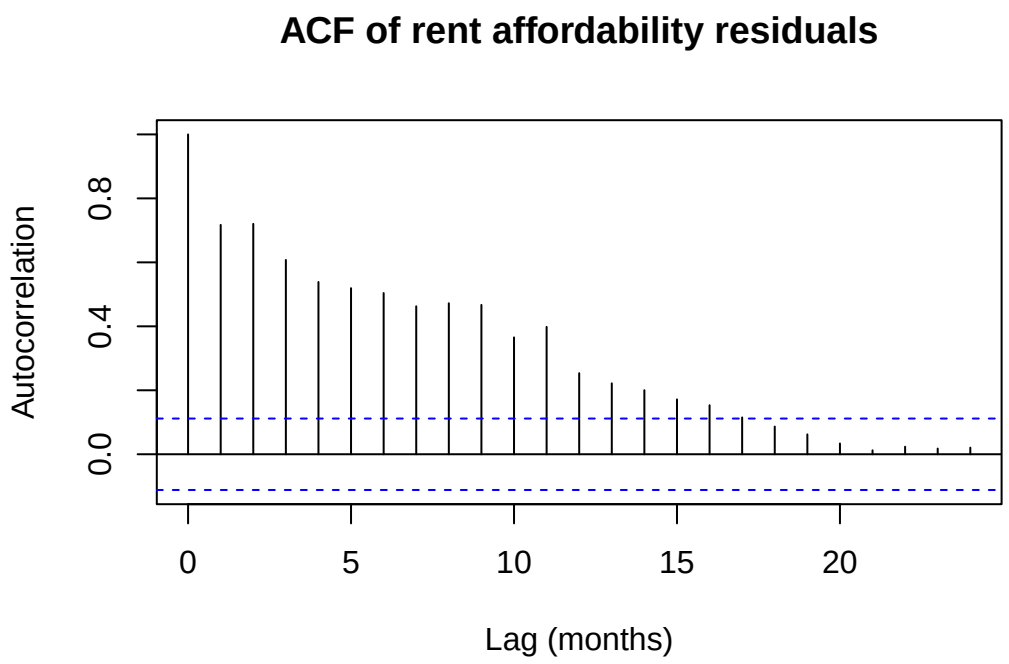



```
par(mfrow = c(1, 1))

# 6.2 Residuals over time
res_rent <- resid(mod_rent)

covid_start <- as.Date("2020-03-01")
covid_end   <- as.Date("2021-12-01")

par(mar = c(4.5, 4.5, 3, 1))

plot(housing$date, res_rent,
     type = "n",
     xlab = "Date",
     ylab = "Residuals (index points)",
     main = "Rent affordability residuals over time",
     bty = "l")

grid(col = "grey90")

# highlight COVID period
usr <- par("usr")
rect(covid_start, usr[3], covid_end, usr[4],
     col = rgb(1, 0, 0, 0.06), border = NA)
```

```r
# emphasize structure first, noise second
lines(housing$date, res_rent,
      col = "grey40", lwd = 1.4)

points(housing$date, res_rent,
       pch = 16, cex = 0.35,
       col = rgb(0.3, 0.3, 0.3, 0.5))

abline(h = 0, lty = 2, col = "grey40")

mtext("Shaded area indicates COVID-19 period",
      side = 3, line = 0.3, adj = 1,
      cex = 0.75, col = "grey40")
```



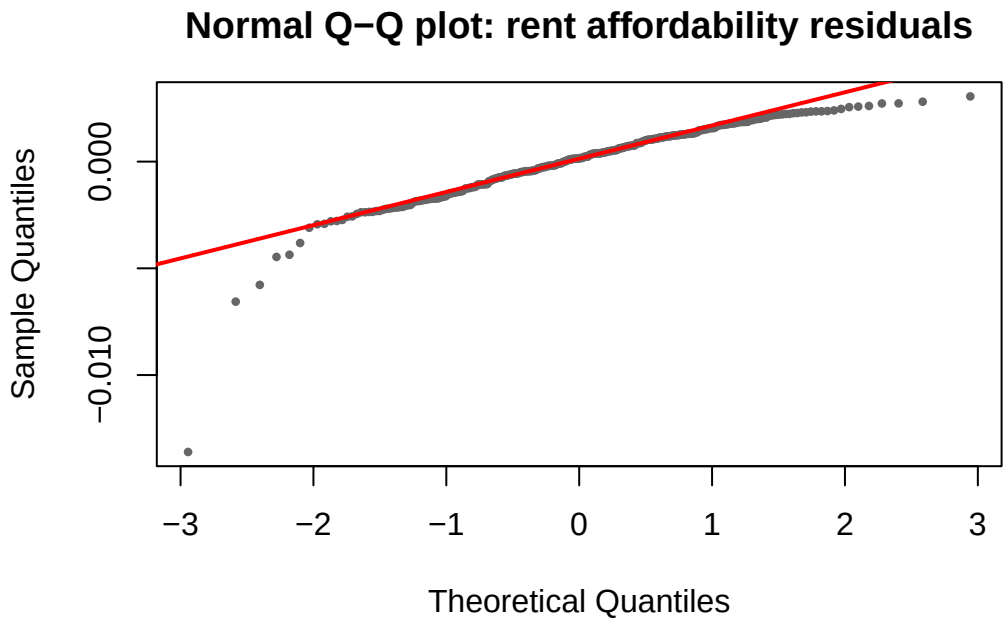

```r
# 6.3 Autocorrelation diagnostics
acf(resid(mod_rent),
    main = "ACF of rent affordability residuals",
    xlab = "Lag (months)",
    ylab = "Autocorrelation")
```




```
# Durbin-Watson test for first-order autocorrelation
# H0: no first-order autocorrelation in errors
dwtest(mod_rent)
```

```
    Durbin-Watson test

data:  mod_rent
DW = 0.56423, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

```
# 6.4 Normality diagnostics
qqnorm(resid(mod_rent),
       main = "Normal Q-Q plot: rent affordability residuals",
       pch = 16, cex = 0.6, col = "grey40")
qqline(resid(mod_rent), col = "red", lwd = 2)
```




```
# Shapiro-Wilk test of normality (large n → very sensitive)
# H0: residuals are normally distributed
shapiro.test(resid(mod_rent))
```

```
    Shapiro-Wilk normality test

data:  resid(mod_rent)
W = 0.88449, p-value = 1.685e-14
```