Give People Housing and They Will Have Kids.
House Prices and Fertility across U.S. Counties, 2000–2023

Filip Zaleski
University of Toronto

April 23rd, 2025

Abstract

This paper investigates the relationship between housing prices and fertility rates across U.S. counties from 2000 to 2023. Using county-level panel data, fixed-effects and first-differenced regressions are used to examine how changes in housing costs shape reproductive decisions. The core finding is that higher house prices are associated with lower fertility, consistent with affordability pressures discouraging family formation and with recent surrounding literature. As a development of this analysis, I explore how this relationship varies across the urban–rural continuum. Results from both fixed-effects and differenced models suggest that the housing–fertility link is stronger in denser counties, indicating that urban affordability constraints are more acute. To strengthen the analysis and add nuance, I implement a regression tree model. I also assess the predictive power of housing sentiment via a random forest approach based on online discourse from Reddit, Twitter, and Google Trends. Sentiment ends up relatively predictive of house prices with a large across-city variation. When interacted with macroeconomic controls, sentiment's role in predicting prices diminishes – reinforcing the primacy of local affordability in shaping demographic outcomes. The paper suggests policy takeaways and points to promising areas of further research.

Contents

1. Introduction

That aging populations pose profound long-term challenges to all global economies has been an axiom in economic literature for decades. An aging population arises naturally from prolonged periods of low fertility, as fewer births gradually reduce the proportion of young individuals relative to the elderly. When occurrent in a society, an aging population burdens public finances through increased pension, healthcare, and social service expenditures (Lee & Mason, 2017). Consequently, understanding and combating the root causes of declining fertility is crucial for sustainable socioeconomic planning in the long term.

Fazio et al. (2024) discovered that gaining access to housing significantly increases fertility outcomes for all adults of childbearing age, but the effect was particularly large for young adults; an increased childbearing probability of 32% and an increase in the number of children by 33%. Motivated by these findings, this paper addresses the central economic question: How do housing prices influence fertility decisions across the United States? As a development of this research question, I also examine whether this relationship differs across the urban–rural continuum. This sub-question has relevance for both federal and municipal policymaking.

The multifaceted nature of households' fertility decisions mean that still today, the causes of low fertility can be hard to identify. Particular to housing, Li (2024) provided global evidence testifying to the relationship between housing prices and fertility over the period 1870 to 2012, demonstrating a long-run relationship between the health of the economy and demographic trends. Dettling and Kearney (2012), in turn, examined metropolitan areas in the United States, finding that there is a notable dependency between rising housing costs and lowering fertility rates. Further, Simon and Tamura (2008) performed a gross historical examination of this correlation in various U.S. cities for the period 1940-2000, once more outlining the link between economic pressure and childbearing decisions. Within a European context, Stoenchev and Hrischeva (2023) specifically investigated the degree to which housing affordability influences family planning decisions in Bulgaria from 2014 to 2021; a contemporary interval. All the aforementioned studies build on foundational economic theory established by Ermisch (1999), who originally described how the decision of young adults to exit parental homes influences their fertility rates.

My paper expands on existing literature by analyzing current and highly pertinent U.S. county data spanning 2000 to 2023, a rich and contemporary focus that is missing from existing literature. Furthermore, I enrich the analysis by including rural areas that have been historically underrepresented in housing-fertility research (exemplified by Dettling and Kearney [2012] amongst others). The study also offers precise input for policymakers seeking

to counter low fertility with more effectively informed housing market policy.

From this point, Section 2.1 provides a detailed explanation of the dataset and data collection processes. Sections 2.2 and 2.3 then offer comprehensive visualizations and summary statistics to provide a preliminary understanding of variable interactions. Section 3.1 provides an overview of the OLS methodology at the heart of this paper, the results of which are reported in sections 3.2 and 3.3. Section 4 adds a machine learning supplement focused on an analysis of online sentiment about the housing market. Finally, Section 5 arrives at a conclusion that summarizes my findings, states their contribution, and pinpoints avenues for future research.

2. Working Data

2.1. Approach and Description

The data set built for this analysis is organized as panel data consisting of more than 12,000 county-level observations in the United States, aggregated per year from 2000 to 2023. This yields a master table with about 212,000 data points. The dependent variable, the fertility rate, was obtained from the Centers for Disease Control and Prevention (CDC) and is defined as the general fertility rate, the live birth rate per 1,000 women of childbearing age (15 to 44).

Housing affordability, the main independent variable, was proxied through the Zillow Home Value Index (ZHVI). It was selected because of its county-level detail across the United States and its relevance in capturing typical home value and affordability. Racial composition was included to control for demographic influences and investigate possible ethnic patterns. Population density and marriage rates, obtained from the U.S. Census Bureau, were included to control for urbanization and social influences on fertility decisions. Unemployment rates from the U.S. Bureau of Labor Statistics and education levels from the U.S. Department of Agriculture were used to describe economic and socioeconomic factors likely to influence fertility.

For sentiment analysis (Section IV), the data was scraped from Twitter, Reddit, and Google Trends using relevant API programs. Control variables for that section include housing demand from Redfin, mortgage rates from the Federal Reserve (FRED), and unemployment from the U.S. Census Bureau.

2.2. Data Summary

Figure 1 contains descriptive statistics of relevant variables. What can instantly be noticed is that the fertility rate is extremely variable averaging approximately 61 births per 1,000 women, but varying widely from 21 to 112. This is as expected for a large and diverse a country like the United States; the variable captures variation in local economic conditions, the cost of housing, and population trends – these all impact fertility. Additionally, the Housing Price Index is also considerably variable with a mean of 231 and a huge standard deviation of 146, indicating considerable housing affordability variation. Large standard deviations in summary statistics are to be expected considering the panel nature of my dataset.

Demographic and socioeconomic characteristics also carry interesting initial insights. Population density averages 902 persons per square mile, though the median of 351 re-

Table 1: Summary Statistics - All Variables

	Variable	Mean	Std	Min	Max	Median
1	Fertility Rate	61	10	21	112	61
2	Housing Price Index	231	146	44	1517	189
3	Population Density	902	2918	7	49536	351
4	Unemployment Rate	6	3	2	29	5
5	% White	82	13	19	98	85
6	% Black	12	12	0	74	8
7	% Indian American	1	3	0	47	1
8	% Asian/Pacific	5	6	0	71	3
9	% Bachelor+	30	10	10	77	29
10	% High School	28	7	7	50	28
11	% Some College	29	5	11	43	30
12	% Below High School	13	7	2	44	12
13	% Married	53	6	31	71	53

veals considerable right-skewness due to very populous urban counties. Unemployment is around 6%, with considerable variation between 2% and nearly 29%, which could indicate regional economic instability and periods of turmoil like the Great Recession or COVID-19 pandemic. The racial composition is predominantly White (82%), with a minority Black (12%), Asian/Pacific Islander (5%), and Native American (1%). Education is fairly dispersed among groups, with around 30% having at least a bachelor's degree. Finally, the marriage rate is 53% on average but with a wide range (31%–71%), which indicates differing family composition across counties. These summary statistics document substantial heterogeneity in the data set and set the stage for subsequent regression analyses in the paper.

2.3. Visualization

Visualizing the data compiled for this research is particularly insightful because of the research's focus on spatial and regional patterns in both fertility and housing dynamics. These nuances might be obscured in strictly numerical summaries. Choropleth visualisations have been constructed and for this purpose and are collected in Appendices 1 and 2. They illustrate county-level absolute changes in the general fertility rate as well as the percentage change in house prices between 2001 and 2023.

Appendix 1 reveals that fertility rates have overwhelmingly decreased across the most populous U.S. counties. Fertility increases, shown in blue, can scarcely be seen. Especially notable is the fact that urban counties - particularly those close to Chicago, Florida, and coastal California - show profound drops, mirroring the emphasis of earlier literature on

urban susceptibility to demographic change. Appendix 2 focuses on the change in house prices during the same period. We can notice a similar urban pattern – coastal, urban counties (California and the Pacific Northwest especially) exhibit huge gains, captured via darker teal shades. In comparison, appreciation in house price in rural counties such as those located in the lower Midwest appears to be relatively modest, though the differences are too subtle to judge visually. We can, through the comparison of the two figures in appendices 1 and 2, infer that regions of high housing price inflation correspond to regions of high declines in fertility rates; further evidence to investigate the hypothesis that urban fertility rates respond more than proportionally to housing market conditions.

3. Econometric Models and Results

3.1. Methodological Approach

This section outlines the econometric strategy used to estimate the key relationship of interest: fertility rates and housing costs in U.S. counties between 2000 and 2023. The analysis is divided across two empirical frameworks: pooled Ordinary Least Squares (OLS) and first-differenced regressions. Both approaches identify variation in fertility outcomes over time, but they differ in how they handle unobserved heterogeneity and in the nature of variation that they capture. The primary body of analysis employs a pooled OLS framework with year and county fixed effects. The former absorb shocks that hit every county in the same calendar year, such as national recessions, federal policy, or COVID-19. The latter absorb all time-invariant differences across counties such as culture, geography, long-run institutions, etc. The primary explanatory variable here is the Zillow Home Value Index, ZHVI. The key control variables were overviewed in section 2. A total of six OLS specifications are estimated. See below for the econometric definition of each model and figure 4 in section 4.2. for results from these models.

(1) $\text{Fertility}_{it} = \alpha + \beta_1 \cdot \text{ZHVI}_{it} + \delta_t + \varepsilon_{it}$

(2) $\text{Fertility}_{it} = \alpha + \beta_1 \cdot \text{ZHVI}_{it} + \boldsymbol{\gamma}'\mathbf{X}_{it} + \delta_t + \varepsilon_{it}$
Where $\mathbf{X}_{it}$ *includes macro controls.*

(3) $\text{Fertility}_{it} = \alpha + \beta_1 \cdot \text{ZHVI}_{it} + \boldsymbol{\gamma}'\mathbf{X}_{it} + \mu_i + \delta_t + \varepsilon_{it}$
Where: μ_i *= county fixed effect,* δ_t *= year fixed effect.*

(4) $\text{Fertility}_{it} = \alpha + \beta_1 \cdot \text{ZHVI}_{it} + \beta_2 \cdot \text{Top30}_i + \beta_3 \cdot (\text{ZHVI}_{it} \times \text{Top30}_i) + \boldsymbol{\gamma}'\mathbf{X}_{it} + \mu_i + \delta_t + \varepsilon_{it}$

(5) $\text{Fertility}_{it} = \alpha + \sum_{q=2}^{5} \theta_q \cdot \text{Quantile}_q + \sum_{q=2}^{5} \phi_q \cdot (\text{ZHVI}_{it} \times \text{Quantile}_q) + \beta_1 \cdot \text{ZHVI}_{it} + \boldsymbol{\gamma}'\mathbf{X}_{it} + \mu_i + \delta_t + \varepsilon_{it}$

(6) $\text{Fertility}_{it} = \alpha + \beta_1 \cdot \text{ZHVI}_{it} + \beta_2 \cdot \log(\text{Density}_{it}) + \beta_3 \cdot (\text{ZHVI}_{it} \times \log(\text{Density}_{it})) + \boldsymbol{\gamma}'\mathbf{X}_{it} + \mu_i + \delta_t + \varepsilon_{it}$

Model (1) begins with a 'baseline' – a simple bivariate regression – while Models (2) and (3) add control variables and county fixed effects respectively, to illustrate the sequential change in coefficients. Models (4) to (6) all incorporate interactions between ZHVI and urbanization that explore how the sensitivity of fertility to housing prices varies by urbanicity, but they differ in measures of capturing urbanization. In (4), urbanization is classified

binarily as top-30%-population-density or not. Model (5) uses quintile-based classification with the first (lowest) quintile omitted. Model (6) uses one variable for population density; a continuous logarithm of population density. This is the preferred specification due to its flexibility and statistical significance. The interaction between log-density and ZHVI in (6) enables an assessment of whether the marginal effect of housing prices on fertility increases or decreases smoothly across the urbanization spectrum. All models are estimated using heteroskedasticity-robust standard errors.

To validate the findings from the fixed-effect models and ensure robustness to time-invariant, county-specific confounders, a second set of models was estimated using first-differenced regressions. This removed all unobserved, time-constant county characteristics by estimating how short-run changes in house prices relate to changes in fertility rates. These three specifications are defined below.

$$(7) \quad \Delta\text{Fertility}_{it} = \alpha + \beta_1 \, \Delta\text{ZHVI}_{it} + \beta_2 \, \text{Top30}_i + \beta_3 \, (\Delta\text{ZHVI}_{it} \times \text{Top30}_i) + \boldsymbol{\gamma}' \, \Delta\mathbf{X}_{it} + \varepsilon_{it}$$

$$(8) \quad \Delta\text{Fertility}_{it} = \alpha + \sum_{q=2}^{5} \theta_q \, \text{Quantile}_q + \sum_{q=2}^{5} \phi_q \, (\Delta\text{ZHVI}_{it} \times \text{Quantile}_q) + \beta_1 \, \Delta\text{ZHVI}_{it} + \boldsymbol{\gamma}' \, \Delta\mathbf{X}_{it} + \varepsilon_{it}$$

$$(9) \quad \Delta\text{Fertility}_{it} = \alpha + \beta_1 \, \Delta\text{ZHVI}_{it} + \beta_2 \, \Delta \log(\text{Density}_{it}) + \beta_3 \, (\Delta\text{ZHVI}_{it} \times \Delta \log(\text{Density}_{it})) + \boldsymbol{\gamma}' \, \Delta\mathbf{X}_{it} + \varepsilon_{it}$$

Notation: Δ denotes first differences; $\mathbf{X}_{it}$ contains the same macro controls as Models (2)–(6), in differenced form (e.g., ΔUnemployment, Δ%Married, Δ%Bachelor's, etc.).

These first-differenced models provide two main advantages. First, they emphasize short-run responses to housing cost shocks. Second, they provide an alternative identification strategy to fixed effects – which are not included in the models. Before running the models, the claim was that if the results from regressions (7) to (9) were to be statistically significant, they would confirm key findings from the first set of regressions – that apart from a significant relationship between housing prices and fertility, the relationship is greater in denser counties. Results are presented in the next section.

3.2. Fixed-Effect OLS Results

Table 1 summarizes the six pooled-OLS specifications with two-way fixed effects and county-clustered standard errors (where $\approx$ 480 counties $\times$ 24 years = 11,611 observations). Moving from model (1) to (3) shows how sequentially adding macro controls and county dummies

notably increases explanatory power (R^2 rises from 0.15 to 0.84) while leaving the ZHVI coefficient stably negative.

Table 2: Fixed-Effects Regression Results for Fertility and Housing Costs

Dependent Variable: Fertility Rate						
	Baseline Regressions			Complex Regressions		
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	65.513***	83.809***	57.752***	66.338***	68.383***	60.512***
	(0.538)	(1.889)	(2.974)	(2.819)	(2.799)	(3.046)
ZHVI	−0.012***	−0.011***	−0.009***	−0.009***	−0.013***	−0.019***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
log(Pop. Density)			1.405***			1.238***
			(0.084)			(0.141)
ZHVI × log(Pop. Density)						0.001***
						(0.000)
Top 30% Density Dummy				3.712***		
				(0.298)		
ZHVI × Top 30% Dummy				0.000		
				(0.001)		
2nd Quantile Dummy					−2.292***	
					(0.450)	
3rd Quantile Dummy					−3.753***	
					(0.443)	
4th Quantile Dummy					0.666	
					(0.460)	
5th Quantile Dummy					2.929***	
					(0.456)	
ZHVI × 2nd Quantile					0.003*	
					(0.002)	
ZHVI × 3rd Quantile					0.006***	
					(0.002)	
ZHVI × 4th Quantile					−0.001	
					(0.002)	
ZHVI × 5th Quantile					0.005***	
					(0.001)	
Observations	11611	11370	11370	11538	11538	11370
R^2	0.150	0.525	0.843	0.846	0.850	0.844
Adjusted R^2	0.148	0.524	0.836	0.840	0.844	0.837
Residual Std. Error	9.603	7.168	4.202	4.167	4.113	4.199
F Statistic	85.077***	380.268***	122.356***	125.072***	127.403***	122.315***
Fixed Effects	*No*	*No*	*Yes*	*Yes*	*Yes*	*Yes*
Controls	*No*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*

Notes: $^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Every iterative complexification of the model from left-to-right adds nuance and approximates the model of interest. Introducing macroeconomic controls in model (2) notably improves explanatory power, while including county fixed effects in (3) controls for time-invariant heterogeneity across counties, which lowers the regression's significance. Regression (4) uses a dummy variable for the top 30% most densely populated counties, along with its interaction with the housing index. Here, the "sensitivity" coefficient directly representing the research question is insignificant, which tells us that the top 30% population density dummy might not be the best measure of urbanization.

Model (6) is the preferred specification because it allows for a fertility–housing gradient that varies continuously with urban density. The relevant coefficients are:

$$\hat{\beta}^{(6)}_{ZHVI} = -0.019 \quad \text{and} \quad \hat{\beta}^{(6)}_{\text{int.}} = +0.001$$

Given the inclusion of county and year dummies, the point estimate implies that – at the sample mean of log(density) ≈ 6.2 – a \$10k (≈ 0.1 SD) increase in house prices is associated with a 0.12-point drop in the general fertility rate. The positive interaction term means that this negative effect attenuates in denser counties – that is, a one-log-unity increase in density (≈ 2.7 times more people per square mile) offsets the ZHVI slope by +0.001, reducing the magnitude of the decline by roughly 5 percent. This may initially seem contradictory to earlier claims that urban areas experience stronger housing-fertility effects. The key distinction is that absolute housing costs are higher in urban counties, so although the marginal effect per \$10k increase attenuates slightly with density, the overall affordability pressure is still more severe in cities. Thus, the combination of high base prices and dense populations still makes urban fertility more sensitive in practice.

3.3. Robustness Check: First-Differenced OLS Results

Table 3 presents three first-differenced regressions that relate changes in fertility to changes in housing costs. Differencing purges all time-invariant country attributes, offering an alternative identification method to the fixed-effects models in section 3.2.

Table 3: Differenced Regressions for Fertility Rate

Dependent Variable: Δ *Fertility*			
	Differenced Regressions		
	(7)	(8)	(9)
Intercept	-0.596^{***}	-0.563^{***}	-0.543^{***}
	(0.078)	(0.089)	(0.077)
ΔZHVI	0.013^{**}	0.014^{**}	0.007^{**}
	(0.001)	(0.002)	(0.001)
Δ log(Pop. Density)			1.103
			(1.971)
ΔZHVI × Δ log(Pop. Density)			0.371^{***}
			(0.063)
Top 30% Density Dummy	0.083^{*}		
	(0.050)		
ΔZHVI × Top 30% Dummy	−0.003		
	(0.002)		
2nd Quantile Dummy		0.007	
		(0.074)	
3rd Quantile Dummy		−0.110	
		(0.073)	
4th Quantile Dummy		0.073	
		(0.073)	
5th Quantile Dummy		0.017	
		(0.074)	
ΔZHVI × 2nd Quantile		−0.004	
		(0.003)	
ΔZHVI × 3rd Quantile		−0.002	
		(0.003)	
ΔZHVI × 4th Quantile		-0.005^{*}	
		(0.003)	
ΔZHVI × 5th Quantile		−0.003	
		(0.003)	
Observations	10824	10824	10824
R^2	0.070	0.071	0.074
Adjusted R^2	0.069	0.069	0.074
Residual Std. Error	2.173	2.173	2.168
F Statistic	74.297^{***}	48.452^{***}	79.099^{***}
Fixed Effects	*No*	*No*	*No*
Controls	*Yes*	*Yes*	*Yes*

Notes: $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

It is important to note that, by nature, first-differenced models remove all between-

county variation - instead relying solely on within-county, year-to-year changes. As a result, the considerably lower R^2 values visible in figure 7 should not be a cause for concern. A series of key insights from table 4 is summarized in table 4 below.

Table 4: Differenced Regression Table Findings

FD Model	**Key Finding**	**Interpretation**
(7) ΔZHVI × Top 30%	ΔZHVI remains positively signed (0.013 **) and significant; the Top 30% dummy itself is positive (0.083 *). The interaction is small and insignificant.	Year-to-year house-price gains are associated with fertility upticks, and those gains are slightly larger in already-dense counties.
(8) ΔZHVI × Density Quintiles	Only the 4th-quantile interaction is (weakly) negative; others are nil.	No clear evidence that incremental changes differ systematically across intermediate density bands.
(9) ΔZHVI × Δ log(Density)	The interaction with Δ log density is large and highly significant (0.371 ***).	Counties that are actively densifying (e.g., via in-migration) experience a stronger positive fertility response to housing-price appreciation.

On top of the peripheral findings reported in table 4, several insights emerge from the first-differenced results - particularly when interacted with fixed effects results from table 2. Firstly, in the level-based fixed effects models (e.g. regression (6)), higher housing prices are associated with lower fertility – following the intuition that unaffordability exerts downward pressure on family formation. In contrast, the differenced models show a positive relationship – year-over-year increases in house prices are associated with increases in fertility. This initially-confusing insight is product of the fact that each model type captures a different time horizon. While the static regressions in (1) to (6) speak to structural affordability, i.e. persistent housing scarcity over a multi-year period, the differenced models (7) to (9) might be picking up transitory wealth effects such as rising housing equity of perceived economic momentum. In other words, short-run increases in housing prices may be interpreted by households as signs of economic strength, thereby increasing confidence in affordability and encouraging family formation—even if, in the long run, high house prices deter fertility. After all, if households interpret rising house prices as a sign of wealth or macroeconomic stability, they may see childbearing as more opportune in the short term.

Secondly, the interaction terms in all three differenced models indicate that the fertility response to housing prices is not spatially uniform. In particular, regression (9) (the differenced alternative of the key fixed effect regression), indicates that densifying counties exhibit

the strongest positive fertility responses to house price appreciation. The reason behind this might lie in migration dynamics: as people move into growing urban areas, it might be the case that the ensuing concentration of economic activity and public support services (both public and profit-driven) encourages childbearing. Alternatively, rising densities may signal broader improvements in local economic conditions that support family growth. Already, the insights are enough to state that both levels of urbanization captured by the first regression set and the changes in urbanization captured by the second set together inform nuances in the housing-fertility relationship.

3.4. ML Supplement: Regression Tree Results

The regression tree analysis performed in this sub-section remains grounded in the structured county-level dataset introduced in sections 2 – 3. Unlike OLS models, regression trees can flexibly capture nonlinear patterns and threshold effects—such as sudden fertility drops beyond a specific education or marriage rate level. Including this method therefore serves as a nonlinear robustness check for the linear econometric framework.

The first model implemented is a restricted tree that uses a simplified set of seven predictors: house prices, unemployment rate, education extremes (percentage bachelor versus under-high school educated), percentage married, and two urbanization dummies (top versus bottom 30% in the population density distribution). Appendix 4 offers a pruned visualization of the restricted regression tree.

The tree selects percentage of bachelor's degree holders as the top split, indicating that education is the primary driver of fertility heterogeneity across counties. This is in line with standard economic theory indicating that education rates are highly related to women's and indeed families' childbearing decisions (Götmark & Andersson, 2020). Lower education (under high school), marriage rates, and to a lesser extent, ZHVI and the top-30% density dummy appear as secondary splits. This sequence of key variables is quite what we would expect - intuitively, these variables represent the set of primary reasons driving household fertility decisions. This model's Mean Squared Error (MSE) value is 67.41, reflecting the average squared difference between predicted versus real fertility values across the test set.

To test the model's extensibility, I trained an unrestricted tree that includes all of 18 independent variables from the full econometric model. Interestingly, this larger model reverses the ordering of the top predictor, where percentage with less than a bachelor's degree becomes paramount. It is then followed by percentage married, the bachelor's degree share, some college attainment, and percentage Black. The unrestricted model achieves a modest MSE improvement of 64.5, representing just a 4.4% gain in predictive accuracy.

However, this comes at a cost of interpretability and parsimony – we more than doubled the number of variables for a small improvement in prediction error. The additional complexity makes it harder to extract economic insight, and risks overfitting given the limited number of counties. The restricted tree offers clearer theoretical intuition.

4. Sentiment Analysis

4.1. The Motivation for Sentiment Modeling

Economic agents do not operate on perfect market information alone; they respond to perceptions, expectations, and wider narratives about future conditions. This behavioural insight is critical in modern economics, particularly in research related to inflation expectations, consumer confidence, and asset bubbles. Roth and Wohlfart (2020) find that subjective expectations are as influential as objective fundamentals in shaping economic behaviour. Fertility decisions, the focus on this paper, hinge heavily on long-term affordability for households and on their perceived stability of the surrounding economy – public expectations about the housing market, for instance, play an important mediating role.

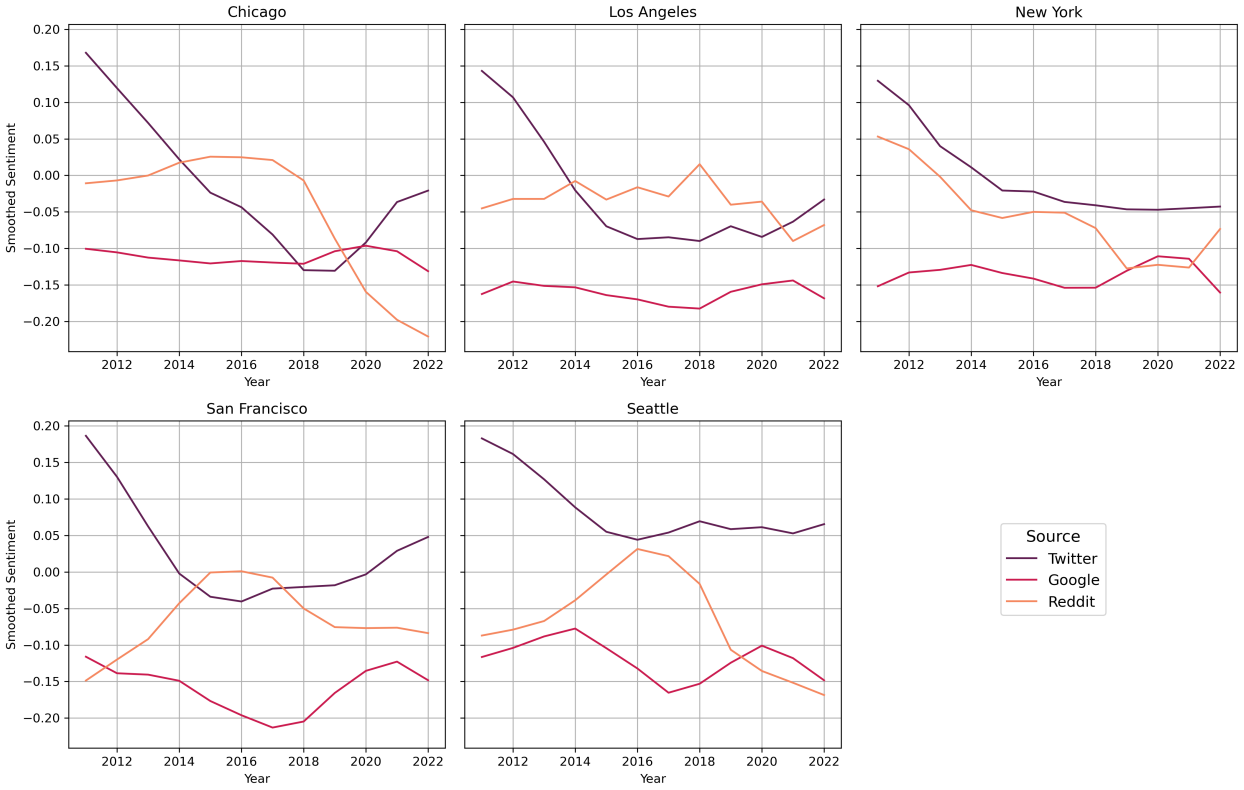
To explore this dimension, I construct a diverse measure of housing market sentiment, aiming to quantify public impressions about affordability and pricing. Sentiment is derived from three platforms:

- **Reddit:** forum-based discussions among largely younger and tech-savvy users.
- **Twitter (now X):** real-time, short-form commentary and reactions from a broad demographic.
- **Google Trends:** aggregate search interest in housing market-related topics.

These sources were chosen for both their breadth and their complementary user bases. While Reddit provides depth and topic-specific focus with deeper and longer threads, Twitter captures immediacy and virality – Tweets are more widespread online than long and consuming Reddit posts. Google Search behaviour, on the other hand, is more of a behavioural signal that reflects widespread and often economically-consequential curiosity. This is a reliable approach for a number of reasons. First, there is existing literature to confirm that sentiment measures derived from social media users closely approximate the sentiment scores derived from more organic, traditional consumer sentiment indicators (Zhang et al., 2024; Fronzetti Colladon et al., 2023).

I restrict my sentiment focus to five U.S. cities; New York, San Francisco, Los Angeles, Chicago, and Seattle. These were selected for their demographic diversity, housing market relevance, and rich volume of online discourse. Figure 1 reveals a graph comparing the smoothed yearly sentiment for from each source across all of the five cities, with time span of 2010 to 2022.

Figure 1: Smoothed Yearly Sentiment Scores by City and Source



A few patterns emerge from the visualisation. First, we can see that sentiment has broadly trended down in all cities across all platforms – consistent with general-knowledge narratives of increasing unaffordability in urban America. Second, Reddit-derived sentiment seems to be the most volatile – perhaps due to its high concentration of housing-related subreddits and event-driven discussions. Third, Twitter sentiment is the most stable and consistently negative. Lastly, Google Trends sentiment shows a unique mid-decade rise in some cities (notably Chicago and Seattle) before declining again after 2018.

These non-rigorous observations suggest that public sentiment surrounding housing has deteriorated over the past decade. Given the potential reflection of these sentiments in people's childbearing choices, the sentiment components is included in the upcoming section's machine learning models.

4.2. Random Forest Results

Random Forests are ensemble tree-based algorithms that perform well in capturing non-linear relationships and high-order interactions among predictors without requiring stringent functional assumptions. They are particularly well-suited for the analysis of sentiment for three reasons. First, the sentiment data came from heterogeneous sources – Reddit, Twitter,

Google – whose effects on house prices may interact in non-obvious and non-additive ways. Second, random forests handle mixed-scale variables (like discrete sentiment mixed with housing prices, etc) and remain robust to multicollinearity among predictors like supply and demand proxies. Third, the algorithm provides intuitive feature importance measures, allowing for the direct comparison of relative explanatory power of sentiment versus economic fundamentals. With these advantages in mind, I estimate two models: a sentiment-only forest (table 5), and an extended forest that adds supply, demand, and financing controls (tables 6 and 7). Table 5 returns the results of a simple random forest model including only the derived sentiment scores.

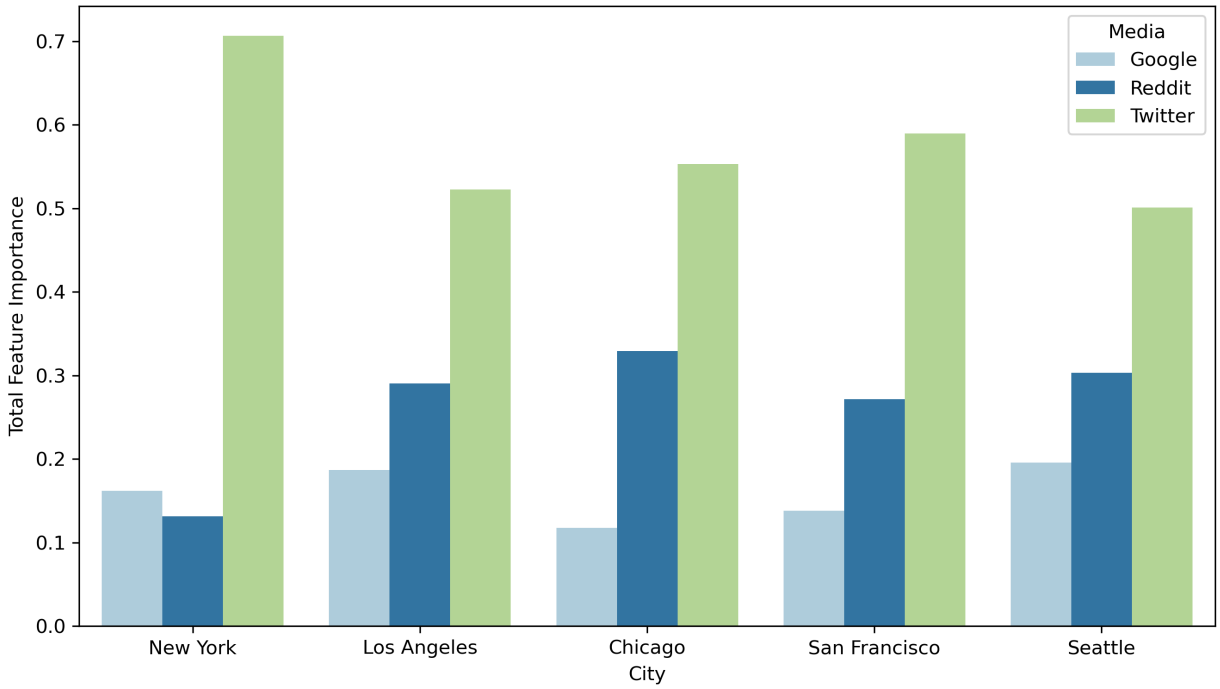
Table 5: Simple Random Forest RMSE Statistics (in Thousands)

Unit: House Price RMSE

City	**Mean**	**Std.**
New York	110	38
Los Angeles	230	104
Chicago	35	12
San Francisco	349	152
Seattle	199	62

Table 5 reveals that model accuracy varies immensely between the five cities. In markets like Chicago and New York, the root mean square error (RMSE) is modest at $35k and $110k respectively. By contrast, San Francisco and Los Angeles have RMSE values in the hundreds of thousands ($349 and $230k respectively). This suggests that for the former, lower-RMSE cities, sentiment scores might be more informative or better-aligned with real house price movements.

Figure 2: Simple Random Forest Feature Importance by Media Type and City



Consider figure 2 for a comparison of the feature importance of each social network sentiment for each city. Since the simple forest included solely different sentiment sources, we can infer the predictiveness of each social network for each of the five cities. Evidently, Twitter dominates across all five cities. There are three reasons for this. First, Twitter simply offers the closest approximation to public sentiment about the housing market out of all of the social networks. Second, the sentiment score constructed for Twitter is the most reliable and accurate – the VADER package was more effective at capturing sentiment from Tweets than FinBERT was for Reddit posts. Third, the substantially larger sample size used for the calculation of Twitter sentiment (versus $\approx$ 70k posts for Twitter versus $\approx$ 12k posts for Reddit) gives a large difference in accuracy. This would be in line with Twitter having the smoothest-shaped lines in the graphs contained in figure 1.

The simple models allows for the conclusion that the predictive power of sentiment alone is limited. Because house prices are determined not just by sentiment but also by fundamentals, I re-estimate the model with controls capturing supply, demand, and financing conditions – the three pillars of house price determination. Housing supply is captured by the logarithm of new listings as well as the logarithm of housing inventory. Housing demand, then, is captured by the unemployment rate and median income. Lastly, the financing of homes is captured by the mortgage rate. Table 6 shows that introducing these variables reduces the RMSE in every city between \$235k (for San Francisco) and \$64k (for New York) on average. This represents a substantial improvement in the model's predictive potential.

Table 6: Controlled Random Forest RMSE Statistics (in Thousands)

Unit: House Price RMSE

City	Mean RMSE	Change (Mean)	Std. RMSE	Change (Std.)
New York	44.5	−65.5	16.6	−21.2
Los Angeles	84.9	−144.7	22.3	−81.4
Chicago	19.6	−15.7	8.3	−3.9
San Francisco	114.2	−234.9	46.3	−106.1
Seattle	74.0	−124.7	17.1	−45.1

We can also deduce that the introduction of control variables has lowered the standard deviation in RMSE within cross-validated folds, implying that the model is not only more accurate, but also more stable. Table 7 displays permutation-based feature importance.

Table 7: Controlled Random Forest Feature Importances

Unit: Percentage Important

City	New Listings	Inventory	Median Income	Mortgage Rate
Chicago	1.6	0.6	68.9	11.4
Los Angeles	0.8	2.6	40.2	9.2
New York	6.1	17.2	69.1	1.9
San Francisco	0.3	1.4	80.0	2.7
Seattle	0.9	0.7	93.8	1.2

City	Google Sent.	Reddit Sent.	Twitter Sent.	Unemployment
Chicago	0.4	1.6	0.5	16.2
Los Angeles	0.2	6.8	0.2	40.1
New York	0.2	0.8	0.3	4.4
San Francisco	0.3	3.7	0.6	11.0
Seattle	0.5	1.1	0.2	1.5

Three dynamics stand out:

1. **Median income dominates in every city (64 to 94% of total importance).** Median income across time is a strong proxy for purchasing power: higher earnings relax borrowing constraints, reduce down-payment challenges, and make monthly mortgage payments more manageable. In supply-constrained markets like San Francisco or Seattle, income becomes almost the sole driver because prices must ultimately align with what households can pay. This aligns with classic affordability theory.

2. **Mortgage rate and unemployment matter, but less so.**
   Mortgage costs explain 2 to 11% of variation, capturing cyclical affordability effects. Unemployment shows a wide range — negligible in Seattle (1%) but highly relevant in Los Angeles (40%) — likely due to differences in local labour-market volatility.

3. **Sentiment is modest but non-zero.**
   The three sentiment scores collectively account for less than 8% of importance in any city — and often less than 2%. This suggests psychological narratives matter only on the margin, overshadowed by economic fundamentals. Still, even a 3–6% predictive contribution may be meaningful in housing markets influenced by consumer psychology, perceived momentum, or narrative contagion. Future improvements to sentiment measurement may enhance its contribution.

The dominance of median income in the random forest underscores a key insight of this paper: affordability, rather than sentiment or spatial categorization alone, is the key driver of housing outcomes. In the OLS models, affordability concerns appear through house price coefficients that are negative in levels (models (1) to (6)) but positive in differenced regressions (models (7) to (9)). The forest-based model builds on this by showing that where income is explicitly included, it trumps all other factors. This reinforces the idea that local purchasing power, not just price levels, constrains or enables movement in the housing market. Meanwhile, the small but nonzero contribution of sentiment supports the notion that psychological and narrative factors are best understood as complements and not substitutes for economic fundamentals. After all, housing is a tangible, dynamic market that is highly researched both by the supply and demand-side.

5. Conclusion

5.1. The Paper Condensed

This research paper has examined how housing prices influence fertility outcomes across U.S. counties – a relationship of growing relevance in both American and global demographic discussions. The OLS approach outlined in section 3.1 allowed me to estimate the marginal effects of key variables under the assumptions of linearity and additivity. It also allowed for the separation of confounding influences like education, unemployment, and marriage rates. Furthermore, the results from my first-differences approach in section 3.3. support the robustness of the paper's main conclusions; that .higher housing prices are negatively associated with fertility. While the marginal effect slightly attenuates in denser counties (due to the interaction term), the overall affordability burden is still greatest in urban areas due to higher base prices. These findings hold even when identification relies only on within-county changes over time, reinforcing the conclusion that housing affordability and urban structure jointly shape reproductive behaviour in modern economies. Nevertheless, OLS as an approach has some limitations that are worth noting. It assumes globally-linear relationships and can struggle to detect nonlinearities or interactions unless explicitly recognized in the model. For this reason, the regression tree framework introduced in section 3.4. requiring no linearity constraints was a valuable addition to my methodology. It accounted for the constraints of the linear methods employed in surrounding literature to uncover nonlinear patterns, interaction effects, and population subgroups that have been overlooked in existing research.

5.2. Policy Recommendations

There are several interesting policy insights that can be drawn from the findings of this paper:

1. **Identifying high-cost areas and expanding supply.**
   Local governments in densely populated counties should consider relaxing zoning and construction restrictions as well as streamlining the approval of new housing projects. Doing so would address one of the root causes of price escalation and help ease the adverse effects on fertility produced by unaffordable housing.

2. **Enhancing affordability through income support.**
   Since income plays a dominant role in moderating the fertility effects of high housing prices, tax credits for young families and down-payment assistance alongside wage

growth — most importantly — may be more effective solutions than direct housing subsidies alone.

3. **Designing geographically-targeted policies.**
   Because the fertility impact of housing costs is stronger in urban counties than rural ones, the scaling of policies such as child allowances should be geographically scaled to reflect local housing market conditions.

4. **Using sentiment as a real-time policy signal.**
   Though I have observed that sentiment cannot act as a substitute for structural data, housing sentiment — especially when monitored in real time (as most macroeconomic indicators cannot be) — may serve as a useful early warning signal for rapid shifts in local housing confidence, which in turn can have detrimental effects on family formation.

5.3. Limitations and Future Directions

The decomposition of my comprehensive sentiment random forest model from section 4.2. into feature importances reveals the limitations of this approach in extracting insights about variable relationships. We can see, for instance, that the model detects some significance from the supply-side of the housing market, but the effects of these (and likely of the remaining) variables may be lessened by the dominance of income that likely captures broad affordability and demand-side dynamics. This illustrates a limitation of using random forests: feature importance reflects marginal contribution within the ensemble, but it does not tell us anything about the isolated or conditional impact of each variable.

Additionally, the limitations of sentiment calculation performed in this paper are worth noting. Firstly, the natural language processing tools used to score Reddit and Twitter posts (like VADER, Flair, and FinBert) remain limited in their ability to capture context, sarcasm, and topic-specific significance. This is something that cannot be significantly improved in the current state of the technological frontier, but future work into word classification and natural language interpretation has the potential to generate an entirely new domain of research in behavioural economics. Secondly, the sentiment dataset used in this paper is small (fewer than 150 thousand city-year observations/posts across all platforms) due to funding constraints, which likely muted its influence in the random forest models. A richer corpus of online content and more advanced language models would likely improve sentiment measurement and help reflect the true relationship between online activity and real-world market dynamics.

6. References

[1] Dettling, L. J., & Kearney, M. S. (2011). House Prices and Birth Rates: The Impact of the Real Estate Market on the Decision to Have a Baby. *National Bureau of Economic Research*. https://doi.org/10.3386/w17485

[2] Ermisch, J. (1999). Prices, Parents, and Young People's Household Formation. *Journal of Urban Economics, 45*(1), 47–71. https://doi.org/10.1006/juec.1998.2083

[3] Fronzetti Colladon, A., Grippa, F., Guardabascio, B., Costante, G., & Ravazzolo, F. (2023). Forecasting consumer confidence through semantic network analysis of online news. *Scientific Reports, 13*(1), 11785. https://doi.org/10.1038/s41598-023-38400-6

[4] Götmark, F., & Andersson, M. (2020). Human fertility in relation to education, economy, religion, contraception, and family planning programs. *BMC Public Health, 20*(1). https://doi.org/10.1186/s12889-020-8331-7

[5] Kharkar, D. (2023, July 9). *About Random Forest Algorithms*. Medium. https://medium.com/@oldshantkharkar3/about-random-forest-algorithms-62163357ab25

[6] Lee, R., & Mason, A. (2017, March). *Cost of Aging – Finance & Development, March 2017*. International Monetary Fund. https://www.imf.org/external/pubs/ft/fandd/2017/03/lee.htm

[7] Li, W. (2024). Do surging house prices discourage fertility? Global evidence, 1870–2012. *Labour Economics, 90*(102572). https://doi.org/10.1016/j.labeco.2024.102572

[8] Roth, C., & Wohlfart, J. (2020). How Do Expectations about the Macroeconomy Affect Personal Expectations and Behavior? *The Review of Economics and Statistics, 102*(4), 731–748.

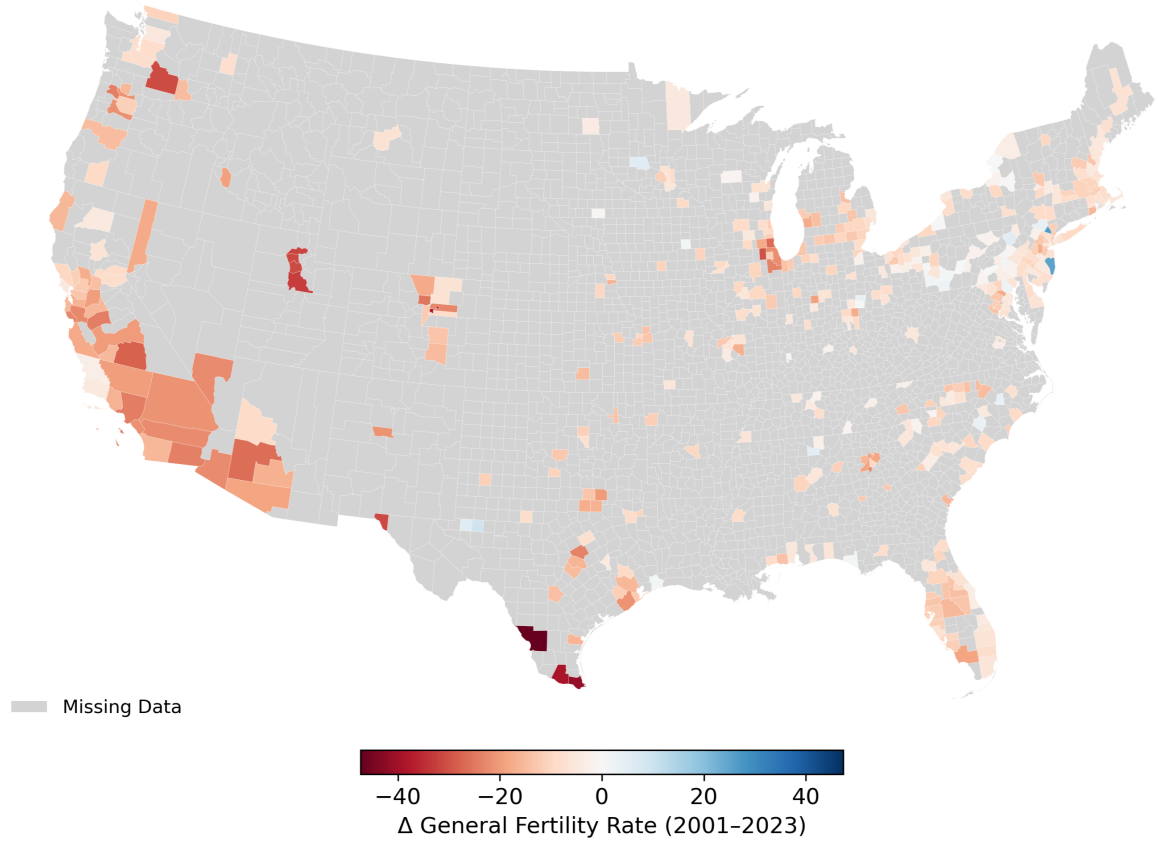
[9] Simcoe, C. J., & Tamura, R. (2008). Do Higher Rents Discourage Fertility? Evidence from US Cities, 1940-2000. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1098847

[10] Stoenchev, N., & Hristoseva, Y. (2023). Study of the Impact of Housing Affordability on the Fertility Rate in Bulgaria (2014–2021): A Regional Aspect. *Baltic Journal of Real Estate Economics and Construction Management, 11*(1), 101–119. https://doi.org/10.2478/bjreecm-2023-0007

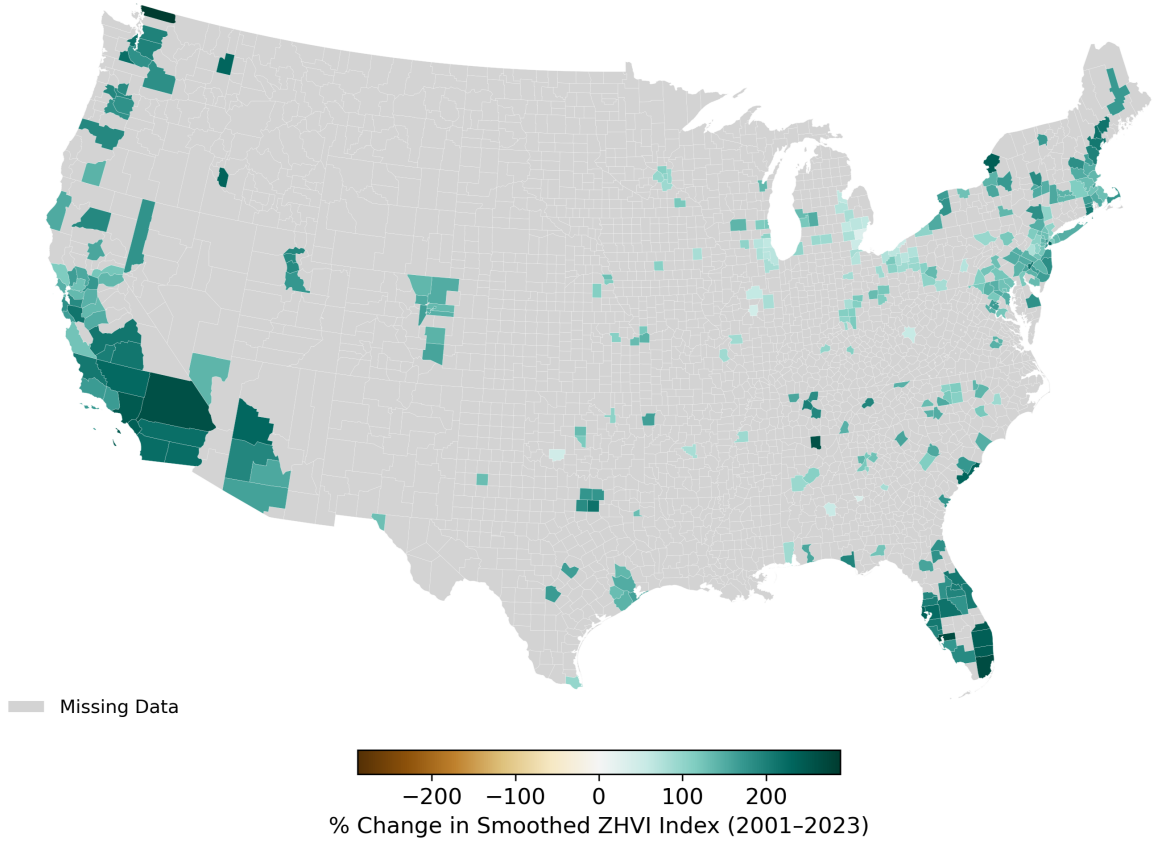
[11] Zhang, Z., Keasey, K., Lambrinoudakis, C., & Mascia, D. V. (2024). Consumer Sentiment: The Influence of Social Media. *Economics Letters, 237*(111638), 111638–111638. https://doi.org/10.1016/j.econlet.2024.111638

7. Appendices

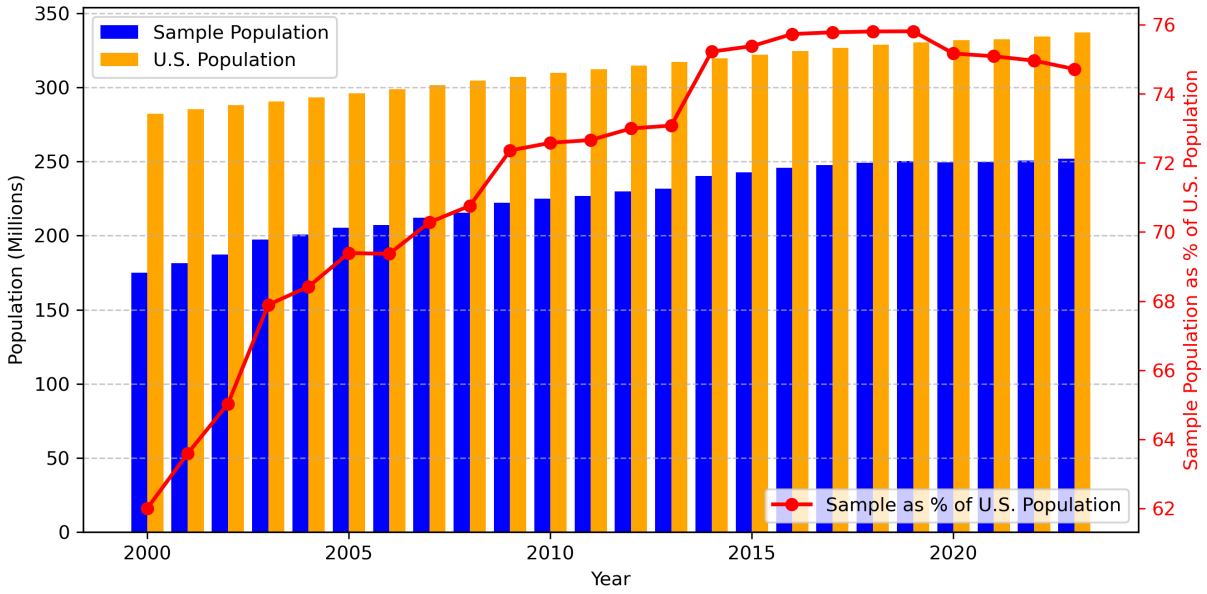
Appendix 1: A Choropleth Visualisation of Fertility Absolute Change Over Time



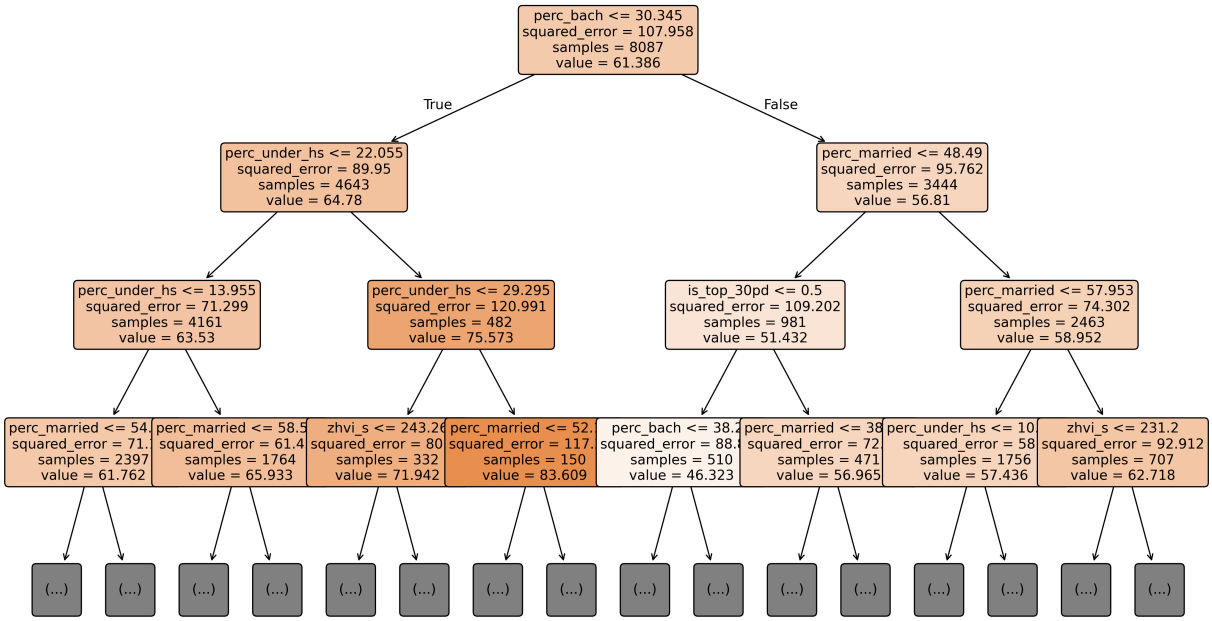
Appendix 2: A Choropleth Visualisation of House Price %-Change Over Time



Appendix 3: Sample Population Versus U.S. Population Over Time



Appendix 4: The Restricted Regression Tree (Main Dataset)



8. Data Sources

The following sources were used to compile data for this research paper:

- Centers for Disease Control and Prevention (CDC). (n.d.). *CDC WONDER*. Retrieved from `https://wonder.cdc.gov/`
- U.S. Census Bureau. (n.d.). *Census Data*. Retrieved from `https://data.census.gov/`
- Federal Reserve Bank of St. Louis. (n.d.). *Federal Reserve Economic Data (FRED)*. Retrieved from `https://fred.stlouisfed.org/`
- U.S. Bureau of Labor Statistics. (n.d.). *Labor Statistics Portal*. Retrieved from `https://www.bls.gov/`
- U.S. Department of Agriculture, Economic Research Service. (n.d.). *County Typology Codes*. Retrieved from `https://data.ers.usda.gov/reports.aspx?ID=4026`
- Zillow. (n.d.). *Zillow Research Data*. Retrieved from `https://www.zillow.com/research/data/`
- Apify. (n.d.). *Tweet Scraper V2*. Retrieved from `https://apify.com/apidojo/tweet-scraper`
- Reddit. (n.d.). *Reddit API Documentation*. Retrieved from `https://www.reddit.com/dev/api/`
- Pytrends. (n.d.). *Pytrends Python Package*. Retrieved from `https://pypi.org/project/pytrends/`
- Redfin. (n.d.). *Redfin Data Center*. Retrieved from `https://www.redfin.com/news/data-center/`