Analyzing Trends in U.S. Residential Real Estate Prices*

A Data-Driven Approach to Understand the Market

Ziheng Zhong

April 17, 2024

This research conducts a detailed analysis of U.S. residential real estate price trends over the past two decades, using data from Zillow to explore influences from geographic location, property characteristics, and economic indicators. By employing multiple regression models, polynomial regression, and Generalized Additive Models (GAM), the study identifies complex interactions that affect housing prices. It highlights a notable shift towards larger homes, potentially driven by changes in work and lifestyle patterns due to recent global events.

Table of contents

*Code and data are available at: https://github.com/iJustinn/House_Price.git

1 Introduction

This paper concentrates on investigating trends in U.S. house prices, a relevant topic as real estate markets globally have undergone considerable increases (Knoll, Schularick, and Steger 2017). Cities such as Toronto, Canada (Courchane and Holmes 2014), and Beijing, China (Deng, Gyourko, and Wu 2012) have seen significant rises in housing costs, indicative of a widespread trend. This study redirects attention to the United States, among the most economically stable and developed nations globally, to ascertain whether similar trends are evident in its real estate market. Through an examination of the U.S. housing market, this research seeks to identify the primary factors that influence house prices.

The U.S. real estate market presents a distinctive case study due to its diverse economic conditions and varied housing markets across states and metropolitan areas (Rapach and Strauss 2009). This analysis will examine variables such as location, number of bedrooms, and other critical factors that are considered to have a significant impact on house pricing dynamics. Understanding these factors is essential, as the findings will offer insights into how policy adjustments and economic changes can influence housing affordability. This research intends to serve as a resource for policymakers, investors, and the public, aiding them in making informed decisions about housing investments and urban planning.

Following this introduction, Section 2 (Methodology), outlines key reasons of the dataset choice, main processing techniques employed, emphasizing transparency and replicability. Section 3 (Results) presents the findings, specifically focusing on the dynamics of house prices across various U.S. states and metropolitan areas. Section 4 (Discussion), analyzes these findings from the perspectives outlined earlier, integrating economic, policy, and regional variables. Finally, Section 5 (Conclusion), summarizes the key insights and implications of this research, offering recommendations for policymakers and stakeholders involved in the housing market.

2 Data

Data used in this paper was cleaned, processed, modeled and tested with the programming language R (R Core Team 2022). Also with support of additional packages in R: `tidyverse` (Wickham et al. 2019), `ggplot2` (Wickham 2016), `janitor` (Firke 2023), `readr` (Wickham, Hester, and Bryan 2023), `knitr` (Xie 2014), `modelsummary` (Arel-Bundock 2023), `testthat` (Wickham 2024), `KableExtra` (Zhu 2024), `viridis` (Garnier et al. 2018), `lubridate` (Grolemund and Wickham 2021), `maps` (Deckmyn et al. 2021), `mgcv` (Wood 2021), `Arrow` (contributors 2024).

2.1 Source

Our study focuses on U.S. house price trends using the extensive datasets available from Zillow Research (Zillow Group, Inc. 2023). This data repository provides a broad array of real estate statistics, capturing the dynamics of house prices across various regions in the United States. The dataset encompasses detailed information on house prices from 2000 to 2024, including breakdowns by state, city, and occasionally neighborhood levels, from median list prices to the Zillow Home Value Index (ZHVI). As a cornerstone of transparency and public engagement, Zillow Research offers open access to these vital real estate data, which are instrumental in understanding housing market behaviors.

We selected this dataset for its reliability, depth, and alignment with our study's objectives. As a trusted source in the real estate market, it provides accurate and timely data essential for our analysis. Its extensive geographic and property-specific details enable an in-depth examination of factors influencing house prices in the U.S. Furthermore, the dataset's structure facilitates efficient analysis and modeling of real estate trends. More details about this dataset can be found in the datasheet in Section A (Appendix).

2.2 Measurement

In this study, the conversion of real-world real estate transactions into quantifiable data entries is meticulously executed using Zillow's extensive datasets. Each real-world transaction is represented as a data point in our analysis, capturing essential attributes such as sale price, property location, and characteristics like the number of bedrooms and bathrooms. For example, a house sale in Miami with three bedrooms and two bathrooms, sold at a specific price, is recorded with precise values for each attribute, including the sale date. This method ensures that each entry in our dataset accurately reflects an actual event, maintaining the integrity and accuracy of the data. The rigorous methodology employed in documenting these transactions allows for precise, data-driven insights into market dynamics, providing a solid foundation for analyzing trends and drawing conclusions about factors influencing real estate prices.

Table 1: Raw Data Preview

RegionID	SizeRank	RegionName	RegionType	StateName	2000-01-31	2000-02-29
102001	0	United States	country	NA	120033.2	120244.4
394913	1	New York, NY	msa	NY	214314.5	215225.3
753899	2	Los Angeles, CA	msa	CA	225004.5	225841.8
394463	3	Chicago, IL	msa	IL	149670.2	149808.8
394514	4	Dallas, TX	msa	TX	125827.2	125883.2
394692	5	Houston, TX	msa	TX	120858.3	120880.8

Table 2: Clean Data Preview

StateName	2000-01-31	2000-02-29	2000-03-31	2000-04-30	2000-05-31	2000-06-30
NA	120033.2	120244.4	120506.3	121068.0	121714.4	122407.5
NY	214314.5	215225.3	216144.4	218007.0	219935.7	222074.0
CA	225004.5	225841.8	226957.1	229176.2	231603.1	234013.0
IL	149670.2	149808.8	150072.7	150729.1	151518.8	152404.9
TX	125827.2	125883.2	125947.7	126115.0	126335.6	126558.9
TX	120858.3	120880.8	120796.5	120846.9	120893.4	121080.6

2.3 Method

Our analysis began with the essential task of refining the raw dataset downloaded from Zillow. The original data was largely retained because it covers a time span that aligns well with the needs of this study. The initial phase involved understanding the dataset's structure. As demonstrated in Table 1, this dataset typically includes multiple attributes such as Region ID, Region Type, location (states), year of transaction, house price, etc. Each record represents a unique property, offering a detailed snapshot of its characteristics.

We then proceeded to clean and process the data as required, setting the stage for more detailed analysis. Our first step involved preprocessing the raw dataset from Zillow, which necessitated several important adjustments to ensure data quality and usability. Initially, we removed the first row of the dataset, which represented national averages rather than state-specific data, to focus our analysis on regional trends. This step was essential as the inclusion of national data could skew the results when examining state-level dynamics.

Next, we streamlined the dataset by removing less relevant columns such as 'SizeRank', 'RegionID', 'RegionType', and 'RegionName'. These columns were extraneous for our analysis, which aimed to concentrate on price fluctuations and trends across different states. By simplifying the dataset as shown in Table 2, we facilitated more focused and faster computations in subsequent steps.

After the initial cleanup, we conducted a more detailed grouping and summarization of the data by state on a monthly basis, as shown in the left table of Table 3. For each state, we

Table 3: Processed Data Preview (by Month & by Year)

StateName	2000-01-31	2000-02-29
AK	133473.93	133669.29
AL	102220.02	102381.62
AR	75198.74	75273.92
AZ	107655.69	107763.46
CA	206610.53	207281.06
CO	171783.37	172140.10

StateName	Year	AvgHousePrice
AK	2000	134747.3
AK	2001	155746.3
AK	2002	170446.5
AK	2003	178964.4
AK	2004	192225.3
AK	2005	219087.3

Table 4: Heatmap Data

region	long	lat	group	order	subregion	AvgHousePrice
alabama	-87.46201	30.38968	1	1	NA	132493.9
alabama	-87.48493	30.37249	1	2	NA	132493.9
alabama	-87.95475	30.24644	1	13	NA	132493.9
alabama	-88.00632	30.24071	1	14	NA	132493.9
alabama	-88.01778	30.25217	1	15	NA	132493.9
alabama	-87.52503	30.37249	1	3	NA	132493.9

calculated the mean house price, excluding any missing values to ensure the accuracy of our results. This aggregation offered a clearer view of the monthly pricing trends across different states, serving as a foundational analysis for identifying patterns and anomalies.

The data was further processed to derive yearly price trends from the monthly data, as shown in the right table of Table 3. We transformed the dataset by pivoting the monthly columns into a single column of house prices, tagged by corresponding months. After converting these month labels into date formats and extracting the year, we calculated the average house price per state for each year. This yearly aggregation aided in examining longer-term trends and making more strategic conclusions about the housing market dynamics.

Finally, as shown in Table 4, to create detailed heatmaps that visualize the distribution of average house prices across the United States, we enriched our dataset with additional geographic information. The subregion column was marked as NA since more detail than cities was not required. This enhancement involved integrating region-specific data, along with precise longitude and latitude coordinates for each state. By incorporating these spatial dimensions, we were able to generate a more nuanced representation of the data, enabling us to pinpoint variations in housing prices with greater accuracy.

By meticulously cleaning and structuring the data through these steps, we ensured that our dataset was not only more manageable but also primed for high-quality, reliable analysis. This thorough preparation was instrumental in supporting our subsequent efforts in charting, statistical analyses, and predictive modeling.

3 Results

3.1 Data Trend

Figure 1 illustrates the average house price trend from 2000 to 2024. This graph highlights several inflection points in housing prices over nearly a quarter of a century. A significant peak is observed in 2006, followed by a notable dip culminating in a trough in 2012. Subsequently, prices ascend, reaching another prominent peak in 2022. The chart clearly marks these fluctuations, emphasizing the cyclic nature of the housing market over time. Excluding these brief highs or lows, the overall trend of housing prices shows a clear upward trajectory.

Figure 2 segments the average house price trends from 2000 to 2024 by house type, ranging from one-bedroom to houses with five or more bedrooms. Each house type displays a distinct trajectory, with the category for homes with five or more bedrooms achieving the highest average prices, particularly showing a sharp rise after 2012. The four-bedroom category also demonstrates substantial growth, though less pronounced than that of the largest homes. One and two-bedroom homes exhibit a more gradual increase, indicating a different progression in price trends compared to larger properties. The disparities among the trends of different house types underscore the varied performance within the housing market. Despite these differences, the overall price trend across different bedroom types consistently shows a clear upward trajectory.

3.2 Heat Maps

The heatmap in Figure 3 illustrates the average house prices across different states in the United States. The color gradient signifies the range of average house prices, with warmer colors (yellows and reds) indicating higher values and cooler colors (purples) representing lower values. States depicted with warmer colors exhibit higher average house prices, while those in cooler colors show lower prices. A noticeable geographical pattern emerges, with certain regions displaying a uniform color, suggesting regional similarities in housing prices. States in the central part of the map typically fall in the mid-range of the color scale, neither the highest nor the lowest in average house price. The variation in color intensity across the states highlights the diversity in average house pricing throughout the country. Additionally, in Figure 3, states on the east and west coasts, such as California and New York, are shaded in the most intense yellowish colors, distinguishing them as having notably higher average house prices compared to states depicted with cooler colors.

Figure 4 consists of a series of heatmaps (labeled a through e) that outline the average house prices by state within the United States, segmented by the number of bedrooms. In the one-bedroom category Figure 4a, the color distribution is relatively uniform, indicating minor variations in average prices across the states. As we progress to the two-bedroom category Figure 4b, there is a noticeable increase in the diversity of colors, suggesting a modest expansion in the price range across different states.

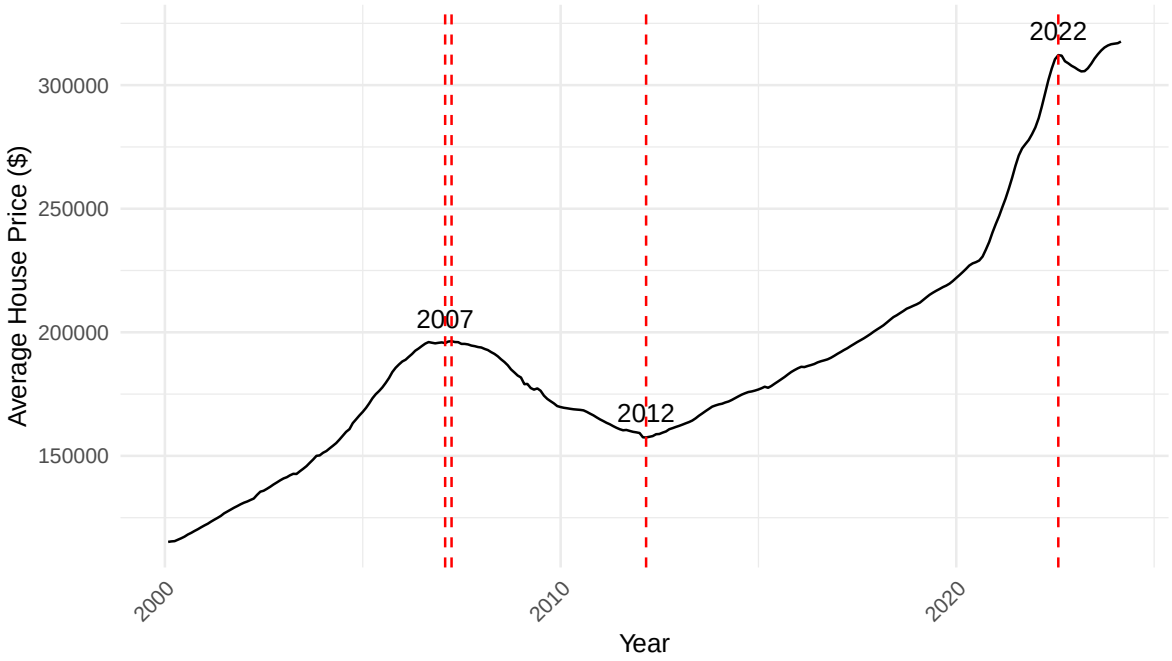


Figure 1: Trend of Average House Price from 2000 to 2024

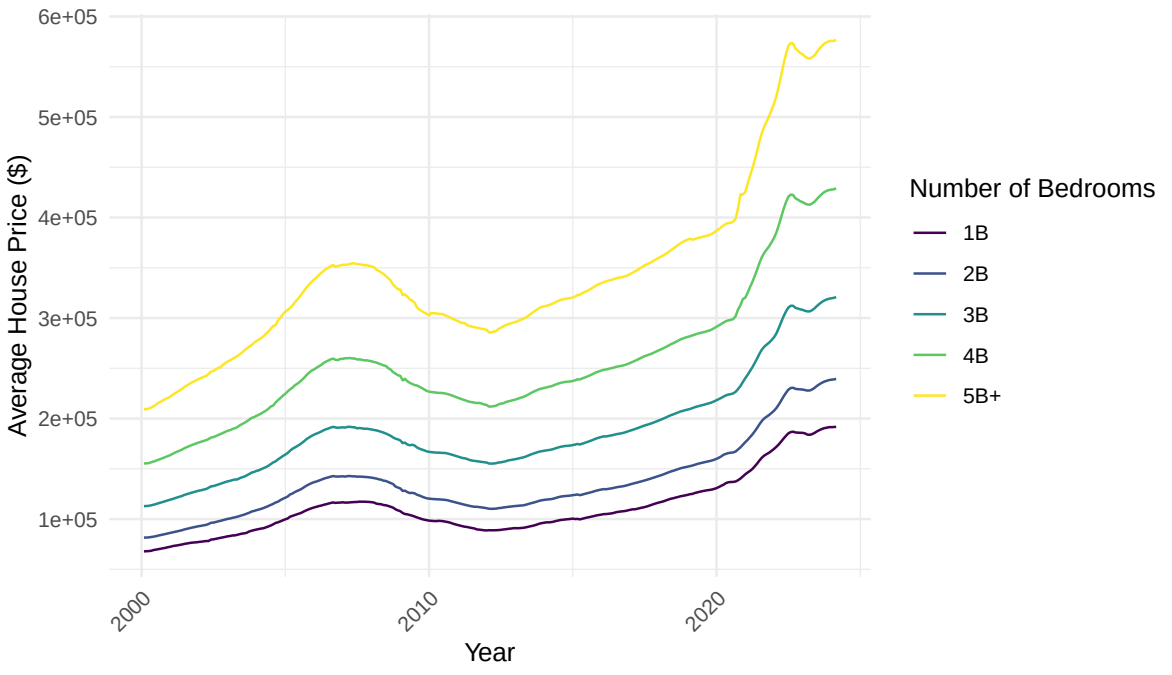


Figure 2: Trend of Average House Price from 2000 to 2024 by House Type

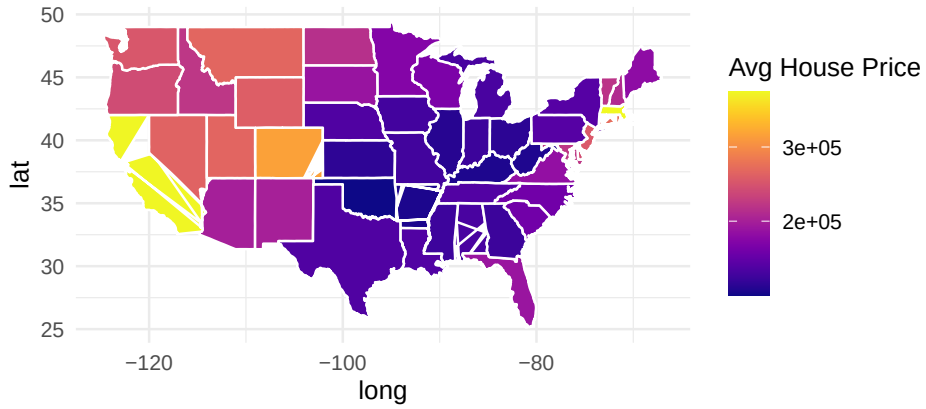


Figure 3: Average Price by State in the US for All House Types

As the number of bedrooms increases to three, as shown in Figure 4c, the color contrast intensifies, indicative of a wider disparity in average prices among the states. This trend continues with the four-bedroom heatmap Figure 4d, where a richer variety of colors emerges. Several states display notably darker hues, signaling higher average prices. This progression highlights the increasing variability in house prices as the size of the properties increases, reflecting differing market dynamics and housing demands across the states.

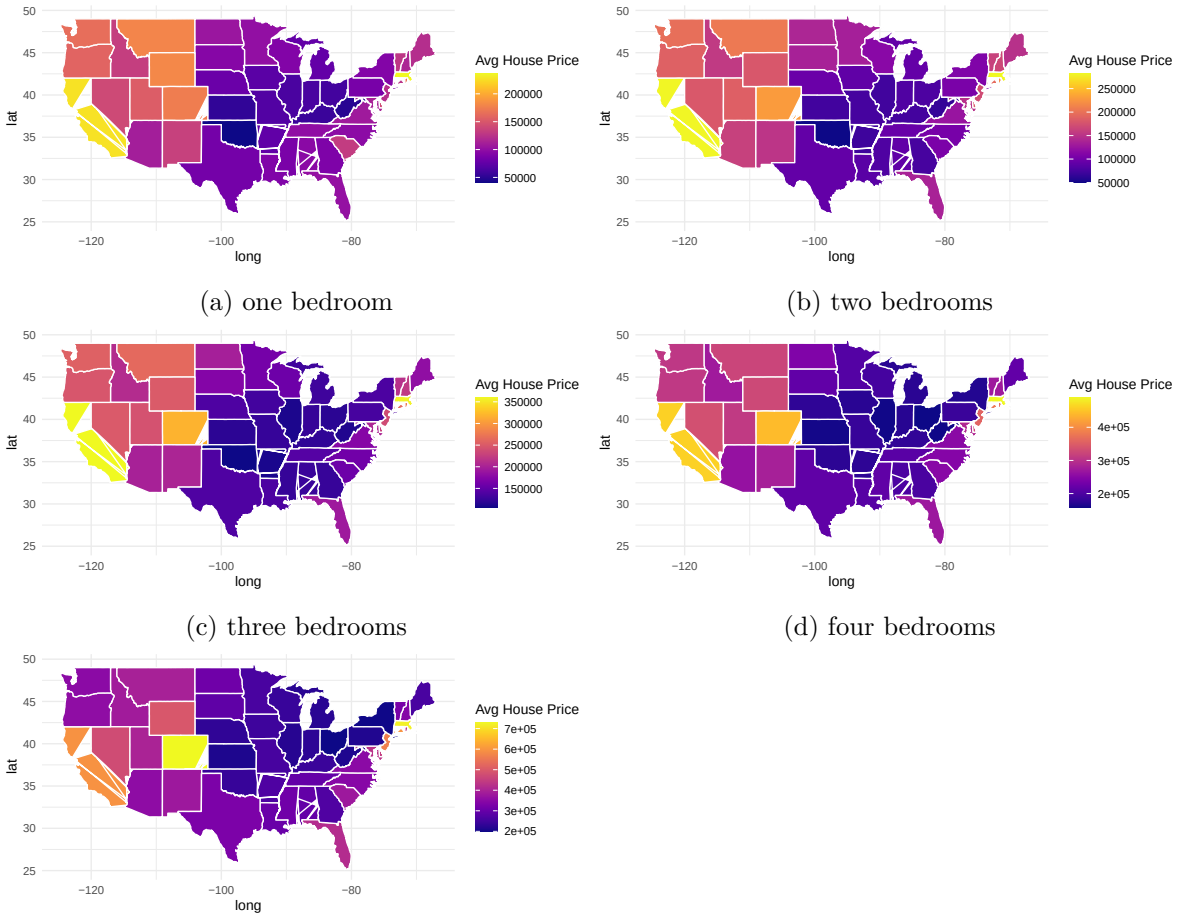
The heatmap for homes with five or more bedrooms, as depicted in Figure 4e, showcases the most striking contrast in color intensity, with several states marked by significantly darker shades. This pattern indicates a broad spectrum of average prices for larger homes, with some states showing considerably higher average prices than others, as evidenced by the more intense coloring. Each heatmap distinctly represents the average price range for each house category, visually emphasizing the spectrum from lower to higher average prices across the states. This visual distinction underscores the significant variation in market conditions for larger properties across different regions.

3.3 Modeling

3.3.1 Model Setup

$$\text{Price} = \beta_0 + \beta_1 \times \text{Year} + \beta_2 \times \text{NumBedroom} + \varepsilon$$

The setup for the Multiple Regression model is straightforward, positing direct linear and additive relationships between the independent variables (Year and NumBedroom) and the dependent variable (house price). This model assumes that changes in the year and the number of bedrooms are directly proportional to changes in house prices, with each independent variable contributing additively to the predicted price.



(a) one bedroom

(b) two bedrooms

(c) three bedrooms

(d) four bedrooms

(e) five plus bedrooms

Figure 4: Average Price by State in the US for Different House Types

$$\text{Price} = \beta_0 + \beta_1 \times \text{Year} + \beta_2 \times \text{Year}^2 + \beta_3 \times \text{NumBedroom} + \beta_4 \times \text{NumBedroom}^2 + \varepsilon$$

The Polynomial Regression model involves a more complex setup, incorporating polynomial transformations of the independent variables to accommodate non-linear trends in the data. Specifically, second-degree polynomial terms are introduced to model curvature, reflecting that the impacts of time and the number of bedrooms on house prices may vary at different rates. This approach recognizes that increases in house prices could either accelerate or decelerate over time, and that the price premium associated with additional bedrooms may not be constant, potentially increasing or decreasing with the number of bedrooms. This model thereby allows for a more nuanced understanding of how these variables interact to influence house prices.

$$\text{Intercept for GAM} = \beta_0$$

The Generalized Additive Model (GAM) provides a flexible approach for modeling complex, non-linear relationships between the dependent and independent variables. Distinct from models that use fixed equations with coefficients, GAM utilizes smoothing functions, enabling it to dynamically adapt to the data without a predefined formula. Additionally, GAM has been shown to fit well in the housing market context (Wang and Chen 2019). In this study, the setup for the GAM involves specifying smooth functions for the predictors, specifically year and number of bedrooms. This approach effectively accommodates non-linear trends that traditional linear models cannot adequately capture.

Unlike polynomial regression, which directly incorporates non-linearity through polynomial terms, GAM utilizes smoothing functions that are fitted to the data in a non-parametric way. This method enables the model to conform to the underlying trend's shape, providing a tailored fit to specific patterns and anomalies in the data. This adaptability makes GAM particularly useful for exploring and understanding the nuanced behaviors of variables within complex datasets.

After gaining some understanding of the model choices, Table 5 and Table 6 display the regression results. These will be explained in detail in later sections. This structured approach helps in comparing the outcomes of different modeling strategies, illustrating how each model handles the complexity of the data and the specific characteristics of the variables involved.

3.3.2 Multiple Regression

$$\widehat{\text{Price}} = -13,969,628.019 + 6,958.769 \times \text{Year} + 59,897.327 \times \text{NumBedroom}$$

The Multiple Regression model provides some intriguing insights into the housing market. It reveals a substantial negative intercept, suggesting that starting from the model's baseline (possibly the earliest year in the dataset), house prices are initially low. The positive coefficient for 'Year' indicates a general trend of increasing house prices over time, with each passing year associated with an average increase in house price of $6,958.769. Additionally, the 'NumBedroom' variable has a positive coefficient of $59,897.327, indicating that houses with more bedrooms tend to command higher prices. This model explains 44.5% of the variability in house prices (R-squared = 0.445), which is respectable but also suggests that over half of the variance is due to factors not included in the model. This highlights the need to consider other variables or more complex models to better capture the dynamics affecting house prices.

Table 5: Modeling Results for Linear Models

	Multiple Regression	Polynomial Regression
(Intercept)	−13 969 628.019	212 976.031
	(395 357.616)	(1359.110)
Year	6958.769	
	(196.466)	
NumBedroom	59 897.327	
	(993.637)	
poly(Year, 2)1		3 866 485.706
		(105 994.679)
poly(Year, 2)2		1 601 816.793
		(105 993.208)
poly(NumBedroom, 2)1		6 591 537.312
		(105 993.313)
poly(NumBedroom, 2)2		1 365 656.076
		(105 994.574)
Num.Obs.	6082	6082
R2	0.445	0.479
R2 Adj.	0.445	0.479
AIC	158 396.4	158 018.2
BIC	158 423.3	158 058.5
Log.Lik.	−79 194.199	−79 003.096
F	2440.891	1397.712
RMSE	109 331.51	105 949.60

3.3.3 Polynomial Regression

$$\widehat{\text{Price}} = 212,976.031 + 3,866,485.706 \times \text{Year} + 1,601,816.793 \times \text{Year}^2$$
$$+6,591,537.312 \times \text{NumBedroom} + 1,365,656.076 \times \text{NumBedroom}^2$$

The Polynomial Regression model reveals notable dynamics within the housing market with its considerably high positive intercept, suggesting that the baseline price for houses at the model's inception is substantial. The incorporation of polynomial terms for 'Year' and 'NumBedroom' illustrates a non-linear relationship between these variables and house prices. Specifically, the positive coefficients for the polynomial terms of 'Year' suggest an initial increase in prices. However, the lower magnitude of the coefficients for the quadratic terms might indicate a slowing growth rate or a potential future downturn in prices.

Similarly, the coefficients for 'NumBedroom' follow a pattern that hints at diminishing returns; the increase in price associated with additional bedrooms becomes less pronounced as the number of bedrooms increases. This model provides a slightly better fit to the data compared to the Multiple Regression model, as evidenced by a higher R-squared value of 0.479. Furthermore, lower AIC and BIC values suggest greater efficiency in model selection, and the reduced RMSE points to improved prediction accuracy. This indicates that the Polynomial Regression model might be more adept at capturing the complexities of the housing market than a simpler linear approach.

3.3.4 Generalized Additive Model (GAM)

$$\text{Intercept for GAM} = 212,976.031$$

The results of the Generalized Additive Model (GAM) as presented in Table 6 show a significant positive intercept, mirroring the polynomial regression model, which suggests a high base price for houses at the outset of the analysis. The R-squared value of 0.506 represents an improvement over the linear models, explaining approximately 50.6% of the variability in house prices. This indicates the GAM's enhanced ability to capture the variance in house prices through its non-linear approach.

The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) values for the GAM are slightly lower than those observed in the polynomial regression model, suggesting that the GAM provides a better model fit with less information loss. However, the marginal reduction in these values implies that while the GAM enhances the fit of the linear models, the improvement in terms of information criteria is not dramatically significant.

The Root Mean Square Error (RMSE) of the GAM is lower than that of the multiple regression but higher than that observed in the polynomial regression, indicating that the model's predictions are reasonably accurate, yet there may be some potential for enhancement. This metric confirms that the GAM has improved the predictive capability of the model, providing an

Table 6

	GAM Regression
(Intercept)	212 976.031
	(1322.680)
Num.Obs.	6082
R2	0.506
AIC	157 697.2
BIC	157 801.7
RMSE	103 028.50

Modeling Results for Non-linear Models

accuracy that is comparable, but not necessarily superior, to the polynomial approach. This performance might reflect the complexities and variabilities inherent in the housing market data, which can challenge even sophisticated models like the GAM.

In Figure 5, the numbers in the parentheses, such as 10.83 and 2.74, represent the smoothing parameters or effective degrees of freedom. These parameters control the flexibility of the spline, balancing the fit of the model against the risk of overfitting by adjusting the "wiggliness" of the curve. The values on the y-axis, which range from negative to positive, quantify the contribution of each smooth term to the predicted value of the response variable, such as house prices. For example, these y-axis values illustrate how changes in the year or the number of bedrooms influence the predicted house price, relative to a certain baseline.

The solid line depicted in the plot is the estimated smooth curve, which shows the modeled relationship between the predictor and the response variable. The dashed lines surrounding the solid line typically represent confidence intervals, providing a range within which the true effect is likely to lie with a certain level of confidence. This visualization helps in understanding the certainty of the model predictions and the potential variability in the effect of each predictor on the response variable.

The plot on the left illustrates the relationship between 'Year' and house prices, revealing a clear non-linear trend. The prices exhibit a cyclical pattern with a notable peak around 2006, followed by a sharp decline, which likely corresponds to the global financial crisis of 2008. After this trough, there is a recovery, with prices trending upwards again, but with indications of leveling off towards the end of the period. This pattern could signify a stabilization in the market or potentially the onset of another downturn.

The plot on the right demonstrates the relationship between 'NumBedroom' and house prices. This relationship is more straightforward: as the number of bedrooms increases, so does the house price. The upward trend appears to be exponential rather than linear, suggesting that additional bedrooms have an increasingly positive effect on house prices. The confidence interval widens with the number of bedrooms, indicating more variability in the data for

houses with more bedrooms. This increased variability could be due to fewer observations in the higher bedroom categories or greater variance in prices for larger homes.

Both plots exemplify the flexibility of GAMs in modeling complex relationships, enabling the identification of trends that simpler linear models might overlook. The confidence intervals demonstrate that the model is fairly confident in its estimates, particularly for the 'NumBedroom' variable. However, the wider intervals at the extremes of the data range suggest a need for cautious interpretation in those areas, reflecting potential uncertainty or the influence of outlier values.

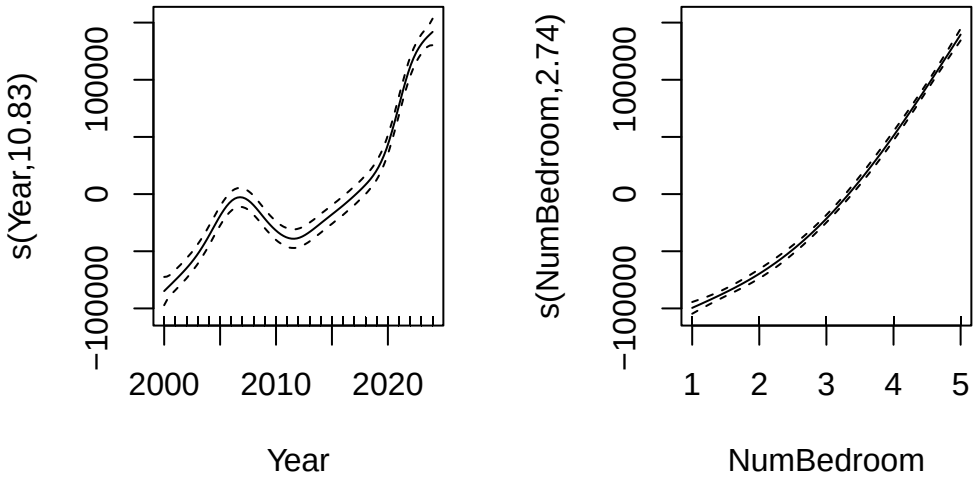


Figure 5

3.3.5 Further Justification

In comparing the Multiple Regression, Polynomial Regression, and Generalized Additive Model (GAM) applied to housing market data, each model serves a distinct purpose based on the complexity of the relationships it can capture between the predictors and house prices.

The Multiple Regression model, with its linear approach, offers a straightforward baseline for comparison. It is the simplest and most transparent, making it an ideal starting point for analysis. However, it is also the most limited in capturing the nuanced behaviors of the housing market. The Polynomial Regression model extends this analysis by introducing non-linearity through polynomial terms, which significantly improves the fit as evidenced by its higher

R-squared value and lower information criteria scores. This model captures more intricate patterns in housing prices related to time and property features.

The GAM further enhances flexibility, not being constrained by predefined polynomial terms, which allows for an even more nuanced understanding of data trends. This model adapts to inherent trends in the data through the use of smoothing functions, leading to an improved R-squared value over the Multiple Regression model. However, the gains in model performance when compared to the Polynomial Regression are modest. While the GAM offers a refined approach, it does not significantly outperform the Polynomial Regression in terms of information criteria and RMSE, suggesting that the increased complexity of the GAM may not be necessary for this particular dataset.

Nevertheless, the visual representation of data trends provided by Figure 5 offers practical insights that might be obscured in tabular regression outputs. For example, the cyclical pattern observed in the 'Year' graph aligns with known economic events that affected housing prices, demonstrating GAM's ability to capture externalities and temporal dynamics beyond what conventional linear models can offer.

The choice between these models should be informed by the balance between the need for simplicity and the desire to capture complex data relationships. While the Polynomial Regression strikes a balance, offering a good fit without overly complicating the model, the GAM provides the most flexibility, which could be crucial for analyzing more intricate datasets or where the precise nature of relationships between variables is less understood.

4 Discussion

4.1 Global Trends

The consistent rise in housing prices documented in this research parallels global economic patterns observed across developed nations (Knoll, Schularick, and Steger 2017). This upward trend is not an isolated phenomenon but part of a broader pattern evident across international borders, highlighting a confluence of macroeconomic and microeconomic factors. This research emphasizes the deep interconnection between the housing market and global economic conditions, illustrating the intrinsic relationship between real estate values and global economic health.

The modeling efforts in this paper underscore the complexity of the housing market's response to economic forces and suggest that straightforward economic models might not always suffice in capturing the real-world dynamics of real estate pricing. The escalation in prices may coincide with inflationary trends affecting various economic sectors, suggesting that housing markets are susceptible to the impacts of global monetary policies and financial climates. These inflationary pressures can distort consumer purchasing power, creating a complex scenario where rising prices may not necessarily reflect actual increases in property value. The data

also suggests that demographic shifts, particularly changes in family structures with more members per household, significantly influence demand and, consequently, prices in the housing market.

Furthermore, the continued increase in house prices in the US has spurred an intriguing area of study. As other research indicates, the GDP of the US has not seen clear growth for an extended period (Antolin-Diaz, Drechsel, and Petrella 2014). It would typically be assumed that without increased earnings, consumer purchasing would not rise; however, the rising house prices contradict this expectation. This discrepancy poses critical questions about the factors driving housing market dynamics, independent of direct economic growth.

4.2 Potential Preference Shifts

The increase in demand for larger homes, as indicated in previous sections, suggests a potential shift in housing preferences, possibly spurred by the recent pandemic, which emphasized the need for multifunctional living spaces that combine comfort with functionality. This shift reflects not only a preference for larger living spaces but also an adaptation to new realities—homes that can accommodate remote work, virtual schooling, and indoor recreation.

This evolving preference could have profound implications, affecting urban planning and infrastructure development. Urban areas, known for high-density living, may need to reconsider zoning laws and development guidelines to meet the growing demand for larger homes. This trend might also influence the architecture and design industries, inspiring innovations in home layouts, building materials, and community planning to cater to new living standards.

Additionally, the preference for larger homes could intensify economic disparities, as the ability to afford such spaces is closely linked to economic status. This dynamic presents a significant challenge for economic planners and policymakers who must ensure that the housing market remains accessible and that housing policies do not exclude those unable to afford larger homes.

Economically, this trend could boost industries related to home construction, renovation, and furnishing as consumers seek to enhance or acquire properties that match their new preferences. However, it also calls for a careful evaluation of the environmental impacts, as larger homes generally require more resources and energy, posing challenges to sustainability goals (Nasir and Colbeck 2013). This balance between accommodating consumer preferences and maintaining sustainability will be crucial for future policy and economic strategies in the housing sector.

4.3 Geographical Position

The geographical analysis of housing price trends, illustrated through heat maps, reveals significant disparities across different regions, depicting an uneven economic landscape. Particu-

larly, coastal regions display elevated housing prices, embodying the enduring principle in real estate that location is paramount. These regions, often centers of economic and cultural activity, have experienced increasing demand due to urban expansion and the attractive notion of coastal living. This demand, coupled with the inherently limited availability of land, leads to a competitive market with steadily rising prices.

The insights derived from the geographical distribution of housing markets underscore opportunities for strategic infrastructural investments. The correct utilization of the Generalized Additive Model (GAM) was instrumental in identifying subtle patterns, such as the impact of microeconomic factors like local zoning laws or the availability of amenities on housing prices. Identifying areas with rapid price growth enables stakeholders to pinpoint locations primed for development or in need of policy adjustments to stabilize the market. Additionally, this data facilitates the targeted allocation of resources to underserved regions, promoting a more balanced economic distribution and enhancing overall economic equity.

4.4 Possible Actions

The insights from this analysis provide a valuable toolset for policymakers, urban planners, and investors, enabling informed decision-making aimed at stabilizing and guiding the housing market. Strategic policy interventions, designed based on the disparities and dynamics revealed by the data, can ensure that the housing market contributes positively to economic stability and community welfare.

For policymakers, the rising housing prices and regional disparities highlight the pressing need for robust housing policies that enhance affordability and accessibility. Government initiatives could include subsidies for first-time homebuyers, incentives for developers to build more affordable housing units, and stricter regulations on speculative real estate investments that inflate prices. Moreover, tax incentives or revisions to zoning laws could promote the development of underused areas, helping to relieve pressure on overheated markets.

The measures implemented in response to these findings must be adaptable and responsive to the shifting patterns of the market. It will be essential to regularly reassess housing policies and investment strategies to respond to new data and changing economic conditions, ensuring that the housing market continues to serve as a pillar of economic stability and a source of equitable opportunities. This proactive approach will help mitigate future market volatilities and foster a robust, inclusive housing market.

4.5 Possible Improvements

The current study provides a detailed analysis, yet there is scope for further refinement. Future research could benefit from incorporating additional variables such as interest rates, housing supply metrics, and construction costs, which could provide a more detailed explanation of price variability. Examining microeconomic factors such as individual income levels and the

impact of local zoning laws could yield deeper insights into market dynamics. Additionally, longitudinal studies that assess the long-term effects of policy changes on housing prices would be highly valuable. Technological advancements in data collection, including real-time tracking of market listings and transactions, could offer more immediate insights into market fluctuations. Expanding the research to include rental markets and the affordable housing sector would enhance the understanding of the broader housing ecosystem.

5 Conclusion

This paper has conducted an in-depth analysis of the U.S. residential real estate market, uncovering the intricate factors that influence housing prices. The study tracked price trends across various states and types of housing, noting significant increases over the past two decades. It demonstrates that house prices are affected not only by traditional factors such as location and property characteristics but also by broader economic indicators and potential shifts in societal preferences.

Our findings indicate that while the U.S. housing market possesses unique attributes, it also aligns with global market trends, especially regarding the rising prices of larger homes. This increase may indicate a shift in consumer preferences, possibly a reaction to the pandemic, towards more spacious living arrangements. Moreover, the geographical analysis reveals a stark disparity in housing costs, with coastal regions showing considerably higher prices, underscoring the critical importance of location in property valuation.

While this study is detailed, there is room for further enhancement and expansion. Future research should incorporate additional economic variables, such as interest rates and housing supply data, to offer a more comprehensive understanding of the market forces involved. Longitudinal studies that examine the effects of policy changes over time could also enrich our understanding of market dynamics. Advances in data collection technology could enable real-time analysis of market trends, allowing for a more responsive approach to rapid changes in the market. Additionally, broadening the scope to include rental markets and affordable housing would provide insights into the entire housing spectrum, crucial for developing holistic housing strategies that benefit all societal segments.

In summary, the U.S. housing market is dynamic and complex, shaped by a wide range of local and global factors. This paper lays the groundwork for a deeper understanding of these factors, offering a fundamental analysis that stakeholders can utilize to make informed decisions that impact community life across the nation.

A Appendix

A.1 Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The datasets are created to support a wide range of real estate analysis and research by Zillow Research group. These datasets typically serve purposes such as market trend analysis, price prediction, housing supply studies, and economic impact assessments. They are designed to fill the gap for comprehensive, accurate, and accessible real estate data for researchers, policymakers, and the general public interested in the housing market dynamics.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The datasets available on the research page of Zillow, a leading real estate and rental marketplace, are typically created by Zillow's own economic research team. This team focuses on providing insights into the housing market and economic trends.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - The datasets on Zillow's research page are generally created and funded internally by Zillow Group, Inc. itself, without specific external grants. As a commercial entity with a vested interest in real estate markets, Zillow utilizes its resources to compile and analyze these datasets for public and internal use.

4. *Any other comments?*

    - None.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - In the dataset used for this paper, each instance representing a geographic region. These regions mainly include the country as a whole (United States), metropolitan statistical areas (e.g., New York, NY; Los Angeles, CA). The dataset contains a series of monthly average house prices, spanning from January 2000 to February 2024, as shown by the date-formatted columns. The data does not mix different

types of instances (like movies, users, and ratings or people and interactions between them) but focuses solely on regional housing price data over time. Each row includes identifiers and names for the regions, the type of region (such as 'country' or 'msa' for metropolitan statistical area), and state names for regions within specific states.

2. *How many instances are there in total (of each type, if appropriate)?*

    - 50 States
    - 895 Regions
    - 290 Dates

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset is likely a sample from a larger set of all possible geographic areas in the United States for which housing data could be collected. This larger set would include a comprehensive collection of all geographic regions in the United States, encompassing every city, town, rural area, and Metropolitan Statistical Area (MSA) with available housing market data. A clear example indicating that this is just a sample is the absence of data for Hawaii. Regarding representativeness, this data is representative, mainly because Zillow is one of the leading real estate trading companies in the United States, and its data covers almost all cities. Additionally, there is no special focus on regions that provide the most insights into the national housing market trends.

4. *What data does each instance consist of? "Raw" data or features? In either case, please provide a description.*

    - RegionID: A unique identifier for each geographic region.

- SizeRank: A rank based on the size or significance of the region, presumably in terms of population or housing market activity.
- RegionName: The name of the region, which can be a country (like the United States), a metropolitan statistical area (MSA), or possibly other region types.
- RegionType: The type of region, such as 'country' or 'msa', indicating the scope or level of the geographic area.
- StateName: The name of the state in which the region is located, applicable to regions within specific states.
- Monthly Average House Prices: A series of columns representing the average house prices for each month, spanning from January 2000 to February 2024. These columns are named by date (e.g., '2000-01-31', '2000-02-29', etc.), with each column representing the average housing price in that region for the given month.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - RegionID: Unique ID associated with each region in RegionName.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - Information such as the latitude and longitude for each region is not included. The absence of geographic coordinates (latitude and longitude) means that while the dataset provides a comprehensive temporal view of housing price trends across different regions, it does not directly offer spatial data that would allow for mapping or spatial analysis of these trends. The reason for missing information is because of the dataset's primary focus on economic trends and prioritizes time-series analysis of housing prices.

7. *Are relationships between individual instances made explicit? If so, please describe how these relationships are made explicit.*

    - The relationships between individual instances (regions) are not explicitly defined in terms of direct interactions or connections between them. The dataset primarily focuses on time-series data for housing prices within various geographic regions without detailing explicit relationships like proximity, hierarchical structures (e.g., how states are composed of multiple cities), or economic interdependencies among these regions.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - There are detailed datasets for houses with specific number of bedrooms. Thus no need to split anything.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - None.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- This is a self-contained dataset.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - None.

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - None.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified.*

    - The data are labled with StateName and RegionName. Therefore, data from each state or region coule be view as a subpopulations of this dataset.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - None.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - None.

16. *Any other comments?*

    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. How were these mechanisms or procedures validated?*

    - The data in the Zillow dataset, specifically the housing prices for various regions, is primarily derived from Zillow's own listings and transactions data, along with public records and assessments. Here's a breakdown of the data acquisition methods:

  - Direct Observation and Public Records: Zillow aggregates housing price data from several sources, including direct listings on their platform, real estate transactions, and public property records such as sales and assessments. This means that much of the data is directly observable or comes from official records, making it a robust and reliable source of housing market information.
  - Zestimate: Some of the housing price data, particularly for times or areas where direct transaction data may be sparse, could be supplemented by Zillow's Zestimate® home values. The Zestimate is an estimated market value calculated using proprietary models that analyze public and user-submitted data. This would fall under data indirectly inferred/derived from other data.
  - Validation and Verification: For directly observed or recorded data (listings, transactions, public records), the validity comes from the data's official or commercial nature. These are factual records of housing sales and listings. For data like the Zestimate, Zillow continuously updates and refines its models based on new data, market trends, and feedback. The accuracy of Zestimates is evaluated by comparing estimated values with actual sale prices when they become available, and Zillow publishes accuracy metrics for its Zestimates, providing a form of validation.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)?*

   - Mainly direct obervation (using data from their platform, as mentioned in previous question).

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

   - Not mentioned by data provider.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

   - Zillow's Platform and Automated Systems: Much of the data collection is carried out through Zillow's own technological infrastructure, which aggregates and processes listings, sales data, and property information from across the United States. This process is automated, involving sophisticated software systems designed to handle large volumes of data.
   - Real Estate Professionals: Realtors and other real estate professionals often use Zillow to list properties. While their primary motivation is to market properties to potential buyers, their contributions add to the dataset's comprehensiveness. Compensation for these professionals comes through the real estate transactions facilitated by their listings, not directly from Zillow for the data per se.

5. *Over what timeframe was the data collected? Does this timeframe match the creation*

timeframe of the data associated with the instances? If not, please describe the timeframe in which the data associated with the instances was created.

- The dataset provided spans from January 2000 to February 2024, indicating that the data collection covers this specific timeframe. This period reflects the creation timeframe of the data associated with each instance, which means each instance's housing price data was collected or estimated for these specific monthly intervals.

6. *Were any ethical review processes conducted? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - For datasets like the one provided by Zillow, which compiles housing market data across various regions in the United States, specific ethical review processes might not be as prominently documented or required in the same manner as they would be for research involving human subjects directly. However, Zillow, like many data-providing entities, operates within a framework of legal and ethical considerations, especially regarding privacy, data accuracy, and transparency.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - Data was obtained from Zillow.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

   - This is a publicly available information on housing prices and transactions.

9. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

   - None

10. *Any other comments?*

    - None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

- Cleaning:
  - Columns (SizeRank, RegionID, RegionType, RegionName) are removed from the dataset, likely because they are not necessary for the subsequent analysis focused on state-level price trends.
- Further Processing:
  - Monthly Aggregation: The data is grouped by StateName, and the mean house price for each state is calculated for each month. This step involves aggregating all numeric columns (assumed to be monthly house price data) and computing their means, excluding missing values.
  - Yearly Aggregation: The monthly data is then transformed to calculate the average house price by year for each state. This involves converting month columns to a long format, extracting the year from each date, and then calculating the yearly mean house price for each state.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

   - The raw data was saved in the project folder [https://github.com/iJustinn/House_Price.git], specific location identified in the README section.

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

   - Programming langugae R [https://cran.r-project.org/] in the IDE RStudio [https://www.rstudio.com/products/rstudio/download/].

4. *Any other comments?*

   - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

   - Building modeling and creating charts for this project.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

   - GitHub Link[https://github.com/iJustinn/House_Price.git]

3. *What (other) tasks could the dataset be used for?*

   - None.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

    - The processed dataset, centered around housing prices across various states and time periods, is limited to only use for similar tasks of this project. Hence it has impact its future use.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - Legal or Regulatory Compliance: Using aggregated and possibly anonymized data for purposes requiring detailed, verified information — such as legal compliance, zoning decisions, or adherence to housing regulations — might not be appropriate. Such applications typically require specific, case-by-case data rather than broad averages or trends.
    - Short-term Investment Decisions: While historical data can highlight trends, using it for short-term real estate investment decisions without considering current market dynamics, economic indicators, and local factors could lead to misguided decisions. The dataset likely does not capture rapid market changes or short-term fluctuations.

6. *Any other comments?*

    - None.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

    - None.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - GitHub, as mentioned in Uses section.

3. *When will the dataset be distributed?*

    - April, 2024. The time when this project is uploaded onto GitHub.

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license*

and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

- None.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

   - None.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

   - None.

7. *Any other comments?*

   - None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

   - The dataset will be held at server of GitHub, regulated by the owner of this project.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

   - The owner can be reached via GitHub account.

3. *Is there an erratum? If so, please provide a link or other access point.*

   - None.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

   - Nothing about this project will updated after it is done.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - None.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

    - Only datasets on the GitHub will be hosted.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

    - People other than the owner can extend the dataset via collaboration feature of GitHub, or directly contact owner for further work.

8. *Any other comments?*

    - None.

References

Antolin-Diaz, Juan, Thomas Drechsel, and Ivan Petrella. 2014. "Following the Trend: Tracking GDP When Long-Run Growth Is Uncertain."

Arel-Bundock, Vincent. 2023. *Modelsummary: Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://vincentarelbundock.github.io/modelsummary/.

contributors, Apache Arrow. 2024. *Arrow: Integration to 'Apache Arrow'*. https://CRAN.R-project.org/package=arrow.

Courchane, Marsha J, and Cynthia Holmes. 2014. "Bubble, Bubble-Is There House Price Trouble–in Canada?" *International Real Estate Review* 17 (1).

Deckmyn, Alex, Original S code by Richard A. Becker, Allan R. Wilks. R version by Ray Brownrigg. Enhancements by Thomas P Minka, and Alex Deckmyn. 2021. *Maps: Draw Geographical Maps*. https://CRAN.R-project.org/package=maps.

Deng, Yongheng, Joseph Gyourko, and Jing Wu. 2012. "Land and House Price Measurement in China." National Bureau of Economic Research.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Garnier, Simon, Noam Ross, Bob Rudis, and Marco Sciaini. 2018. *Viridis: Default Color Maps from 'Matplotlib'*. https://CRAN.R-project.org/package=viridis.

Grolemund, Garrett, and Hadley Wickham. 2021. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Knoll, Katharina, Moritz Schularick, and Thomas Steger. 2017. "No Price Like Home: Global House Prices, 1870–2012." *American Economic Review* 107 (2): 331–53.

Nasir, Zaheer Ahmad, and Ian Colbeck. 2013. "Particulate Pollution in Different Housing Types in a UK Suburban Location." *Science of the Total Environment* 445: 165–76.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rapach, David E, and Jack K Strauss. 2009. "Differences in Housing Price Forecastability Across US States." *International Journal of Forecasting* 25 (2): 351–72.

Wang, Pei-De, and Mingchin Chen. 2019. "THE NON-LINEAR RELATIONSHIPS OF NUMERIC FACTORS ON HOUSING PRICES BY USING GAM." *Journal of Data Science* 17 (1).

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2024. *Testthat: Get Started with Testing*. https://CRAN.R-project.org/package=testthat.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wood, Simon N. 2021. *Mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. https://CRAN.R-project.org/package=mgcv.

Xie, Yihui. 2014. *Knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

Zillow Group, Inc. 2023. "Zillow Research Data: Home Values." https://www.zillow.com/research/data/.