Econometrics

Determinants of House Prices in Texas

Name: Akhilendran Indrabalan

Student ID: 0338860, 2073641, 0326771

Teacher: Dr. Indika

Course Code: BUAN-B360-001

Due Date: December 13, 2024

Introduction

The real estate market stands out as one of the significant determinants of the regional economy, as well as the wealth of an average individual and investment decisions. In Texas, the increase in population and urbanization has increased the demand for residential houses, making it a fertile ground of study. However, the factors that determine the value of real estate are intricate and range from individual property attributes to other properties and macroeconomic factors. These comprehensions are essential for homebuyers to investors, and policymakers who want to make rational decisions.

This research focuses on determining the factors that affect the list price for properties in Texas by analyzing data collected from 500 property listings. The intention is to evaluate the influence of the property's square footage, the number of bedrooms and bathrooms, and the type of property on list prices. Furthermore, city and postal code-specific analyses will be conducted to explain regional trends in the specific area. The results will provide an understanding of the housing affordability issues, assist investors in searching for the target markets, and assist real estate agents in evaluating real estate.

To understand the factors affecting housing prices, the following literature review surveys several studies examining these factors using various methodologies. The literature can be categorized into three major themes: income and affordability, urbanization, and macroeconomic and regulatory factors, and microeconomics and property characteristics. (Saiz, 2024)

Looking into the literature from the income and affordability approach, various findings supported the positive relationship between income growth and housing prices. According to Gyourko, Mayer, and Sinai, rising household incomes strengthen purchasing power. Thereby increasing demand for housing and increasing prices. This suggests that housing markets are sensitive to changes in income population levels. Greater income will lead to a heightened demand for housing. (Gyourko; Mayer; & Sinia, 2006)

In contrast, other studies argue that the elasticity of the housing supply can relieve the impact of income growth on prices. In markets where the housing supply is more sensitive, the price increase due to higher income is less pronounced if the region’s infrastructure cannot accommodate it properly. (Malpezzi, 1999).

Another perspective is looking at real estate pricing through Urbanization; population growth is cited in multiple sources as a vital driver of housing demand. Multiple researchers indicate that areas with fast population growth experience a significant increase in housing prices. Especially when faced with land limitations, such as mountains and coastline. This can play a crucial role in housing affordability in growing urban areas.

The impact of urbanization on prices is more pronounced in cities with geographic constraints, where rapid population growth increases housing price inflation due to land limitations. Furthermore, land restriction can completely impact the effects of population growth. Certain studies support the idea that regulations due to land limitation result in higher income prices by limiting new housing developments in desirable locations. (Glaeser, Gyorko, & Saks, 2005)

Studies in macroeconomics and regulatory factors have examined the relationship between economic indicators and the CRE market. For example, Glancy and Kurtzman focused on interest rates and loan performance in assessing the crisis in the CRE sector. Furthermore, the IMF explained the consequences of commercial real estate concerning the economy's stability, particularly the exposure of banks to CRE assets. (Glancy, D.; & Kurtzman, R., 2024)

Multiple studies support the role of microeconomics and property characteristics in real estate pricing models, including factors such as property characteristics, location, and size in square footage. After identifying household income and property size as key determinants of housing prices (Gyourko; Mayer; and Sinai, 2006). Other studies have highlighted the importance of location with properties in urban areas that need different property characteristics depending on family size or necessities, increasing the demand.

The housing market is influenced by individual property features, buyers' preferences, and the decisions of households. Microeconomic factors focus on the behavior of buyers willing to pay depending on their preferences. As we noticed, individual income levels and property attributes shape the demand for housing. These factors impact how much buyers are willing to pay for specific properties, such as size, number of bedrooms, and location. (Gyourko, J.; Mayer, C.; & Sinai, T., 2006)

The following regression analysis objective is to understand residential real estate pricing in Texas using property characteristics. This analysis tries to determine and measure the relationship between the characteristics of properties, for example, the size of the house in square feet, the number of bedrooms and bathrooms, the size of the lot, the age of the property, and location factors and how these factors affect the housing prices. It uses statistical methods to isolate the most influential factors determining the value of properties to gain a better insight into the Texas real estate market. Such findings may be extremely useful for different players in the market, including buyers, sellers, real estate agents, and policymakers, given that this market is quite dynamic. The study's objective is to create a predictive model to improve decision-making and offer useful information concerning residential real estate appreciation in Texas.

Data and Methodology

This study implements a detailed econometric model to determine which factors affect the residential property listing prices in Texas using a sample of 500 property listings from Texas Real Estate Trends 2024 Kaggle dataset (Kaggle, 2024). However, it should be noted, that the sample only reflects a part of the market, and it is treated as sufficient representative under the Central Limit Theorem allowing the statistical observations derived to estimated larger market behaviors and trends. By examining both structural and geographical aspects alongside with transformations of the key variables we aim to identify which factors have the most significant impact on the listing prices.

The main dependent variable considered in our analysis is the List Price of each property, measured in USD. The independent variables capture a broad range of factors that may influence List Price. These include both structural characteristics such as Type of the Property, Square Footage of the property, number of Bedrooms, number of full Bathrooms, number of Stories, and the Year Built of the property—and geographical variables, such as City and Postal Code.

To gain new insights and interpret underlying patterns, we used several transformations. For example, we took the natural logarithm of selected variables—number of full bathrooms, bedrooms, square footages, stories, year built to mitigate issues of skewness and to interpret changes in property prices proportionally. Then, to explain the non-linear and categorical effects on housing pricing, we have also included quadratic logarithm transformations of square footage and year built as well as a dummy variable of type of property.

Based on our observations, it seems that single-family houses make up for about 92% of the listings alone. Other property types, such as mobile homes, farms, and condos, appear less frequently but remain relevant to niche markets. Most listing prices range between $300,000 and $500,000, with a few reaching up to $28.95 million which indicates us skew toward higher-end properties.

The market trends suggest that there is a drastic change in the consumer behavior with the buyers wanting to purchase homes that are smaller and more compact starting, in the years of 2023-2024. This gradual change in preference along with the increase in new buildings after the year of 2000 due to the boost in the economy and increased population indicates that not only contemporary style but other demographic changes are influencing the resale price of housing in Texas (Kanchana, 2023). Such findings point towards the fact that square footage along with the number of bedrooms, bathrooms and property age could be likely determinants of listing prices.

To better comprehend these trends, we utilized various visualization methods. The histogram offers insights into the distribution of listing prices. Bar charts highlight differences in property types and a

Geomap to provide a spatial perspective, allowing us to identify areas with distinct pricing patterns and characteristics.

For this analysis, we decided to apply a multiple linear regression model to estimate the effect of the identified independent variables on the listing prices. Considering structural variables, quadratic and interaction terms and categorical indicators help us to conduct a comprehensive analysis of how property characteristics interact to influence list price.

Null Hypothesis (H_0): At least one structural characteristic such as square footage, number of bedrooms, number of bathrooms, and year built do not have a significant effect on listing price.

By using a solid econometric model, applying appropriate data transformations, and conducting careful exploratory analysis, we aim to understand the key factors that influence residential property listing prices in Texas.

Table 1: Summary Statistics

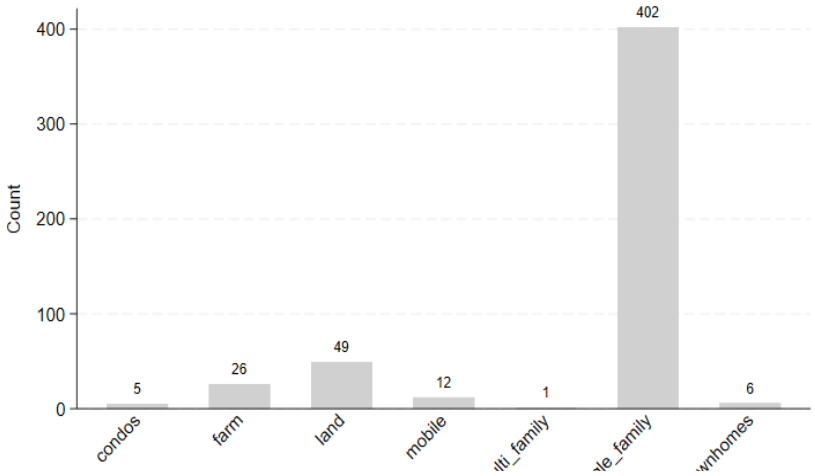
	(1)	(2)	(3)	(4)	(5)
VARIABLES	N	mean	sd	min	max
Price	501	510,669	1.318e+06	10,000	2.895e+07
Bathrooms	436	2.333	0.759	1	8
Beds	440	3.455	0.914	0	9
Square_footage	438	2,335	3,220	0	67,139
Floors	391	1.376	0.526	1	4
Year_built	289	2,000	25.40	1,891	2,024
Property_type	501	5.411	1.314	1	7

There are 501 observations of listing prices, with a mean of $510,669, a standard deviation of 1.318, indicating positive skewness, a minimum of $10,000, and a maximum of $28,950,000, indicating high variability. Full bathrooms (baths_full) have a mean of 2.333 (SD = 0.759), ranging from 1 to 8, while bedrooms (beds) average 3.455 (SD = 0.914), ranging from 0 to 9. Square footage (sqft) has a mean of 2,335 (SD = 3,220), with a minimum of 0 and a maximum of 67,139. Stories average 1.376 (SD = 0.526), ranging from 1 to 4, and year built averages 2,000 (SD = 25.40), spanning from 1891 to 2024. The variable (type_new), representing property types, has a mean of 5.411 (SD = 1.314), ranging from 1 to 7.

Figure 1. Data distribution using k-density of List Prices

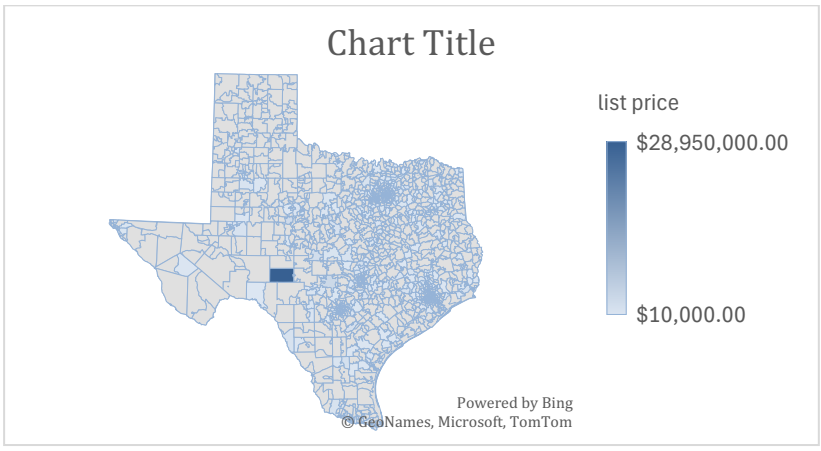
Note: Prices are skewed right, with most properties priced between $300,000 and $500,000.

Figure 2. Bar Chart of Property Types



Note: Single-family homes dominate the market, representing 92% of all listings.

Figure 3. Geographical Distribution of Texas Listing Prices



Note: The map reveals distinct clusters of higher-priced properties within Texas.

Model

We looked to prior studies before defining the models in an attempt to identify the key factors influencing home prices. Numerous studies have shown structural characteristics such as the number of bathrooms and full bedrooms, the square foot area, and the age of the property. For example, in the study on the Atakoy housing market, the authors claim that factors like the area of the house, the number of bathrooms, and the category of the dwelling unit have a bearing on the market prices (Şahin, 2023).

In all the models the dependent variable is the natural log of housing prices ($lnlistprice$). The natural log transformations can also make the interpretation of coefficients easier by treating them as elasticities while also avoiding problems related to variance. Key explanatory variables in natural log are number of full bathrooms, number of bedrooms, square footage of the property, and the year built. To account for non-linear effects, we included quadratic terms. To evaluate the combined effects and account for variations among property categories, we used dummy variables for property types (such as farm, single-family, and townhome) and interaction terms.

Equation

$$ln(listprice_i) = \beta_0 + \sum_{k=1}^{4} \beta_k \mathrm{X}_{ki} + \sum_{m=1}^{7} \gamma_m \mathrm{Z}_{mi} + \varepsilon_i$$

Note: The model includes structural variables, interaction and quadratic terms, and dummy variables for property types, which show category effects on prices but are not written in the equation for simplicity.

X_{ki} = $ln(baths\_full_i)$, $ln(beds_i)$, $ln(sqft_i)$, $ln(year\_built_i)$
Z_{mi} = Interaction and quadratic terms $ln(sqft) \times ln(baths\_full), ln(sqft)^2$, etc
$\mathrm{type\_new}_{ni}$ = dummy variables for type of the property

This equation was created using information gathered from previous research. The important role of structural characteristics such as property size and age in determining home prices (Şahin, 2023). Similarly, a study demonstrated on how the inclusion of quadratic terms enhances our comprehension of pricing dynamics for residential housing markets (Zietz, 2008).

The regression model used in this study operates under key econometric assumptions to ensure the accuracy and the validity of its findings. The assumption of linearity states that there is a direct linear relationship between the dependent variable, the natural logarithm of list price and the independent variables such as the property characteristics. This assumption allows us to interpret the estimated coefficients as direct effects on the dependent variable.

Another crucial assumption is homoscedasticity, which requires that the variance of the error term (ϵ) remains constant across all levels of the independent variables. This helps to ensure the reliability of significance tests and confidence intervals. Finally, the model assumes normality of error term implying that residuals are normally distributed, which is essential for reliable statistical inference, particularly when doing the hypothesis testing.

This model's objective is to divide the variables that affect list prices into measurable parts to provide useful information. By integrating a combination of log transformation, quadratic terms, and dummy variables to account for categorical differences, this model provides a robust framework to explore variations in housing prices across Texas. The use of logarithmic transformations is used to handle potential challenges with skewed data and provide proportionate interpretations. By accounting for non-linear influences like decreasing returns on the effect of square footage or year built, quadratic terms further improve the model.

Results

Table 1 reports the regression results analyzing factors influencing real estate prices in Texas. Model (4) was selected as the best fit for addressing the hypotheses due to its inclusion of log-transformed variables, interaction terms, and quadratic transformations, which provide a nuanced understanding of how property characteristics affect list prices.

By using double log method, the coefficients are percentage changes in the list price for a 1% change in predictor variables, holding all other factors constant. The independent variables were determined to be statistically significant by means of p-value approach where any p-value of less than 0.05 indicates the null hypothesis can be rejected.

The log of full bathrooms ($lnbaths\_full$) has a coefficient of 0.428, hence, according to the model, for a 1% increase in the number of bathrooms is associated with a 0.428% increase in the list price. This variable is very significant with p = 0.001 and this further strengthens bathroom features as a determinant property value.

The log of the number of bedrooms (ln_beds) has a coefficient of 0.677, which means that an increase in the number of bedrooms positively increases the list price. However, this effect is not significant at 0.152 level of significance. Also, the quadratic term for bedrooms (ln_beds^2) has coefficient -0.461 (p = 0.064 implying that the more bedrooms there are, the lower the effect there is. We did further examination of these coefficients, and it revealed that the maximum value where the effect of the addition of bedrooms

is still beneficial is 2 bedrooms. This turning point was calculated with the formula (−b/2a) and then the log value of 0.735 was transformed into its original world figure. After this point, the marginal impact of including more bedrooms in the property has begun to decrease which means, when compared to other features in the house, having extra bedrooms may not be highly preferred by the buyers.

Additionally, it was noted, the interaction term for the square footage with property type (ln_sqft_type) has a coefficient of 0.185 ($p < 0.01$). This means that for every 1% increase in the square footage, makes the list price to increase by 0.185% but this effect also depends on the type of property. This result emphasizes the difference in the effect property size has on property price among different property categories.

The log of year built ($ln\_yearbuilt$) has a negative coefficient of -4142.75 ($p = 0.044$), which indicates that relatively newer properties have a lower price, holding other factors constant. However, the quadratic term ($ln\_yearbuilt^2$) with a positive coefficient of 273.23 ($p = 0.043$). This suggests that moderately aged properties may in the beginning lose their value, newer homes eventually regain a premium maybe because of the demand for more modern amenities and construction. The unusual size of the coefficients probably indicates issues with the scaling of variables and characteristics of the dataset itself. Specifically, the year built of the property ranges from 1891 to 2024, this also results in wide numerical variation. This range when subjected to log transformation, leads to differences in values and enlarges the coefficient estimates. In some cases, it could also represent segments of the market that prefer older homes in older neighborhoods, as well as understand that they are willing to pay a premium for new homes.

Dummy variables for property type provide further insight into how property categories influence pricing compared to the baseline group. Compared to the baseline property type mobile homes, single-family homes, and townhouses, are expected to result in a significant reduction in list price. Farms are associated with a -107.7% decrease ($p = 0.001$), Mobile homes with a -486.3% decrease ($p < 0.01$), Single-family homes with a -728.7% decrease ($p < 0.01$), and Townhomes with a -899.9% decrease ($p < 0.01$). These coefficients are relatively high since they are compared against a baseline category that likely represents a niche, high-value property type, which emphasizes the relative differences in pricing.

Overall, we observed that the area of the property, the facilities offered like bathrooms and year built of the property was built tend to affect the pricing the most. The interaction between square footage and the type of property explains how buyers' preferences for extra area, to a certain extent, are influenced by the property type, which in turn indicates a variety of buyer preferences in the market. Meanwhile, year

built of the property and bedroom count have non-linear relationships that show complexity in property valuation. For full bathrooms, the null hypothesis is rejected, which indicates their role in estimating properties is quite substantial. For year of built, the null hypothesis is rejected since both the linear and the quadratic terms are significant and slight aged houses lose their value while new emerging ones gain a premium, but this has a cap. In contrast, the null hypothesis for bedrooms failed to reject its linear term is non-significant but the quadratic term proves the existence of diminishing returns. These results reflect the dynamics that exist in the housing market in Texas in relation to property features and consumer preferences.

Table 2: Determinants of real estate price in Texas in 2024.

VARIABLES	(1) Model 1	(2) Model 2	(3) Model 3	(4) Model 4	(5) Model 5
Log of Bathrooms	-0.394	0.313	4.789	0.428***	176.4**
	(1.704)	(0.226)	(3.935)	(0.128)	(72.66)
Log of Bedrooms	0.960	0.296	0.616	0.677	-1.259
	(1.938)	(0.375)	(2.532)	(0.472)	(2.747)
Log of Square Feet	-13.212	1.255	-10.04		-100.3
	(46.064)	(1.850)	(46.36)		(61.05)
Log of Year Built	-7.053	-4,183***	-4,335***	-4,143**	-4,568***
	(45.885)	(1,575)	(1,588)	(2,044)	(1,576)
Log of Square Footage x Bathrooms	0.176		-0.648		-0.503
	(0.242)		(0.569)		(0.567)
Log of Square Footage x Bedrooms	-0.140		-0.0426		0.250
	(0.269)		(0.374)		(0.403)
Log of Square Feet x Year Built	1.878		0.973		13.03
	(6.059)		(6.040)		(8.011)
Log of Bathrooms x Bedrooms	-0.399				-0.820*
	(0.330)				(0.474)
i.Property_type = 0, Farm	0.396	0.330	0.338	-1.077***	0.372
	(0.251)	(0.244)	(0.249)	(0.318)	(0.248)
i.Property_type = 1, Mobile Home	-0.497**	-0.541**	-0.563**	-4.863***	-0.503**
	(0.245)	(0.239)	(0.242)	(0.438)	(0.242)
i.Property_type = 2, Single Family	-0.192	-0.268	-0.251	-7.287***	-0.165
	(0.212)	(0.206)	(0.208)	(0.645)	(0.211)
i.Property_type = 3, Townhomes	-0.588**	-0.627**	-0.617**	-9.000***	-0.566**
	(0.259)	(0.255)	(0.256)	(0.829)	(0.256)
Log of Square Feet^2		-0.0129	0.277		0.160
		(0.122)	(0.283)		(0.285)
Log of Year Built^2		275.9***	285.4***	273.2**	295.9***
		(103.7)	(104.7)	(134.5)	(103.8)
Log of Bathrooms^2		0.0771	0.395		0.732**
		(0.143)	(0.318)		(0.346)
Log of Bedrooms^2		-0.315*	-0.311	-0.461*	-0.190
		(0.172)	(0.240)	(0.248)	(0.266)
Log Feet x Property Type^2				0.185***	
				(0.0159)	
Bathrooms x Year Built^2					-22.67**
					(9.582)
Constant	58.366	15,860***	16,475***	15,714**	17,638***
	(348.839)	(5,979)	(6,025)	(7,763)	(5,988)
Observations	283	283	283	283	283
R-squared	0.624	0.634	0.636	0.645	0.645

Notes: The dependent variable is the natural log of the list price (lnlistprice. Model 4 includes predictors for log-transformed bathrooms bedrooms year built and square footage adjusted for property type. Interaction and quadratic terms are added to capture non-linear effects. Diagnostic check multicollinearity issues: ln(year_built) and ln(year_built^2), The VIF exceeded 725,000. This is typical when a variable and its squared term are both included, these variables were retained because of their importance in capturing non-linear relationships (Zietz, 2019). Property type is treated as a categorical variable. Diagnostic checks confirm no omitted variables (p = 0.1721 and no heteroskedasticity (p = 0.6953. N = 283. Adjusted R-squared = 0.632. Standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Conclusion

The data collected in this study gives a detailed and concise analysis of the factors affecting Texas residential real estate prices. The analysis of 500 property listings established that property attributes like the size of square feet, number of bedrooms and bathrooms, and the year of construction are the most important factors determining a property's price. Out of these, the size of the house in terms of square footage was found to be the most crucial factor that influenced property value, which is in line with prior research conducted in this regard. The inclusion of quadratic terms showed that property size is a declining function, highlighting the fact that bigger houses are more valuable, but the value that is added by increasing the size of the house at some point becomes minimal.

The correlation between the year built and the low price of older properties revealed a possible bias that could be considered counter intuitive. The quadratic terms for new locations and properties showed that recently built homes had higher return rates. This implies an added value for modern homes and well-equipped properties in certain areas. Also, the type of property played a major role in identifying the pricing dynamics since most of the listings were for single-family homes, and the prices for such properties varied greatly from other types of properties, including mobile homes and townhomes.

From a practical point of view, these results present relevant conclusions for investors, real estate agents, and policymakers. Investors will be able to understand that properties with larger square footage and located in established neighborhoods can help generate high returns. while real estate agents may leverage insights about structural and locational characteristics to set competitive listing prices. On the other hand, policymakers may use these findings to address housing affordability issues by identifying areas where property characteristics influence prices and creating strategies to ensure fair access to housing.

Despite the findings of this study, which provide a strong econometric base and many useful conclusions, there are also some suggestions for future research. The multicollinearity issue detected in some variables, including the year built and its quadratic term, calls for developing more sophisticated models or collecting more data to capture the non-linear relationships. Enlarging the dataset containing properties and including macroeconomic variables that could enhance the analysis will be useful. This research adds to the existing literature on housing markets by providing a theoretical and applied framework for understanding and predicting changes in the Texas real estate market.

Comments to Feedback from professor and Referee Report

The feedback and corrections received following our rough draft submission very helpful in understanding our data and formatting guidelines for the project. Fixing formatting errors, we started off by creating a title for the cover page. Next, we went through each header and removed numeric listing, only focusing on bold heading titles for reach section. We combined the sections: Literature review, and its subcategories to all be included in the projects introduction, making it easier to understand and more streamline. Next, we combined the Methodology sections with Data Summary and Visualization as well as its subcategories again, to instead make a Data and Methodology section. This included correcting bolding errors, adding a geo-map using excel, and correcting the null hypothesis. For the summary statistics and regression output, we renamed variables so all readers can understand our methodology. Under the Model section, we first corrected the variables names just created, then removed our multicollinearity table, and instead moved the explanation to the regression notes. We removed all unnecessary equations and focused only on our model, correcting its format. The results section was rewritten to only use p-value as the general approach when interpreting regression results, as well as each variables significance. Sign, Size, and Significance was also added under the Model section to explain our variables importance and how it supports our results. As cited above, the regression output was corrected with normal variable names, corrected format, and a clearer section. Lastly, we ordered our references from A to Z to follow correct formatting. Overall, the suggestions were immensely useful for us and helped us create a better project.

Our team is aware of the importance of feedback and comments as tools to enhance the quality of our regression analysis. Feedback is crucial in the refinement of our approach and in enhancing the accuracy of the study as well as the overall robustness of the study. We went through all the suggestions made in the three pieces of feedback to determine the areas that could be enhanced with the help of the suggestions. All these were considered and incorporated into our analysis to fill gaps, enhance the method, and produce more reliable and valuable outcomes. That is why, accepting this feedback, we have increased the rigor and depth of the analysis and ensured that our conclusions meet the highest standards of quality and relevance. Our effort will help to incorporate all helpful feedback to make our work more reliable and effective.

The first report helped us to be more concise when printing the results. This feedback pushes us to improve how we transmit the results and the analysis applications. “The results section gives numbers but doesn’t explain what they mean in the real world.” In the final submission, we tried to change how we transmitted data to be more reader friendly. Changing the results entirely and being more conscious of the interpretation. The second feedback report, “Make sure that you just stick with your research on residential real estate and do not include commercial real estate. Make sure to remove this section since it differs from

residential real estate and does not directly." Thank you for this feedback. We will be careful to address the information and not expand to areas not covered in the analysis.

References:

- Case, K. E., & Shiller, R. J. (2003). Is there a bubble in the housing market? Brookings Papers on Economic Activity, 2003(2), 299–362. Available at: https://www.brookings.edu/bpea-articles/is-there-a-bubble-in-the-housing-market.
- Fan, G. (2006, November). Determinants of house price: A decision tree approach. ResearchGate. Available at: https://www.researchgate.net/publication/238398459.
- Fendoglu, M. S. (2021). Commercial real estate and financial stability: Evidence from the US banking sector. International Monetary Fund. Available at: https://www.imf.org.
- Glancy, D., & Kurtzman, R. (2024). Determinants of recent CRE distress: Implications for the banking sector. Finance and Economics Discussion Series 2024-072. Washington: Board of Governors of the Federal Reserve System. Available at: https://www.federalreserve.gov.
- Glaeser, E. L. (2005, May 1). Why have housing prices gone up? American Economic Review. Available at: https://www.aeaweb.org/articles?id=10.1257/000282805774669961.
- Glaeser, E. L., & Ward, B. A. (2009). The causes and consequences of land use regulation: Evidence from Greater Boston. Journal of Urban Economics, 65(3), 265–278. Available at: https://www.sciencedirect.com/science/article/pii/S0094119008001043.
- Gyourko, J., Mayer, C., & Sinai, T. (2006). Superstar cities. National Bureau of Economic Research. Available at: https://www.nber.org/papers/w12355.
- Hilber, C. A. L., & Robert-Nicoud, F. (2013). On the origins of land use regulations: Theory and evidence from US metro areas. Journal of Urban Economics, 75, 29–43. Available at: https://www.sciencedirect.com/science/article/pii/S0094119012000570.
- Hwang, M., & Quigley, J. M. (2006). Economic fundamentals in local housing markets: Evidence from U.S. metropolitan regions. Journal of Regional Science, 46(3), 425–453. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9787.2006.00423.x.
- Jafari, A. (2019, June). Driving forces for the US residential housing price: A predictive analysis. ResearchGate. Available at: https://www.researchgate.net/publication/333855171.
- Jude, G. D. (2002, February). The dynamics of metropolitan housing prices. ResearchGate. Available at: https://www.researchgate.net/publication/5142158.
- Kaggle. (2024). Texas Real Estate Trends 2024 dataset. Available at: https://www.kaggle.com.
- Kanchana. (2023). Texas Real Estate Trends. Retrieved from https://www.kaggle.com/code/kanchana1990/texas-real-estate-trends.
- Malpezzi, S. (1996). Housing prices, externalities, and regulation in U.S. metropolitan areas. JSTOR. Available at: https://www.jstor.org/stable/24832860.
- Malpezzi, S. (1996, July). Economic analysis of housing markets in developing and transition economies. ResearchGate. Available at: https://www.researchgate.net/publication/23741915.
- Quigley, J. M., & Raphael, S. (2005). Regulation and the high cost of housing in California. American Economic Review, 95(2), 323–328. Available at: https://www.aeaweb.org/articles?id=10.1257/000282805775014236.
- Saiz, A. (2010, October). The geographic determinants of housing supply. ResearchGate. Available at: https://www.researchgate.net/publication/227347270.

- Şahin, E. (2023, February). Factors affecting housing prices: The case of Istanbul-Atakoy. ResearchGate. Available at: https://www.researchgate.net/publication/373957018.
- Zietz, J. (2008, February). Determinants of house prices: A quantile regression approach. ResearchGate. Available at: https://www.researchgate.net/publication/23534659.

Appendix: Do file

```
*playing with the data for some insights
clear all
cd "C:\Users\aindraba\Documents\final_project\datasets"
use "real_estate_nov5.dta", clear
reg list_price baths_full_calc sqft type year_built

// Generate a new variable from 'location' by splitting and keeping relevant parts
gen clean_location = regexs(1) if regexm(location, "([^_]+)")

// Replace dashes with spaces for better readability
replace clean_location = subinstr(clean_location, "-", " ", .)

// Display the first few cleaned entries to verify
list location clean_location in 1/10

**********************************************************************
**Cleaning and Analysis of Real_estate_Nov% (most updated)**
**********************************************************************
clear all
cd "C:\Users\aindraba\Documents\final_project\datasets"
use "real_estate_nov5.dta", clear

list in 1/10 //Displaying the first few rows

**Fill missing listPrice values with the mean:
summarize list_price, meanonly
replace list_price = r(mean) if missing(list_price)

drop if missing(baths_full) | missing(beds) // Dropping rows where baths_full or beds are missing
drop if missing(beds) | missing(sqft)

generate price_per_sqft = list_price / sqft //generate price_per_sqft

summarize list_price sqft beds baths year_built //Summary Statistics for Key Variables:

pwcorr list_price sqft beds baths year_built, star(0.05) //Correlation Matrix

* Clean histogram with customized y-axis formatting to avoid scientific notation
histogram list_price, normal kdensity bin(30) ///
    title("Distribution of Listing Prices", size(medium)) ///
    xlabel(0(500000)3000000, format(%10.0gc)) ///
    ylabel(, format(%10.0gc) nogrid angle(0)) ///
    xtitle("Listing Price ($)", size(medium)) ///
    ytitle("Frequency", size(medium)) ///
    graphregion(color(white)) ///
    plotregion(margin(zero)) ///
    legend(off)



* Summary statistics for the 'list_price' variable
summarize list_price

* Additional detailed percentiles (like 25%, 50%, 75%)
centile list_price, centile(25 50 75)


* Create a horizontal bar graph of property type counts
graph bar (count), over(type, label(angle(45))) ///
    bar(1, color(gs13)) ///
    ylabel(, angle(0) format(%10.0gc)) ///
```

```
    ytitle("Count") ///
    blabel(bar, format(%10.0gc)) ///
    graphregion(color(white))

                tabulate type


****************************************          **************************
*******************************NOV 18*********************************
************************************************************************

reg list_price baths baths_full beds sqft stories year_built price_per_sqft

**Generate Type as numerical variable
encode type, gen (type_new)

reg list_price baths baths_full beds sqft stories year_built price_per_sqft type_new

*****
** Data Cleaning
******
**Starting over to
// Step 1: Load your dataset
clear all
cd "C:\Users\aindraba\Documents\final_project\datasets"
use "real_estate_nov18.dta"

* Calculate the mean of listPrice (excluding missing values)
summarize listprice, meanonly

* Replace missing values in listPrice with the mean
replace listprice = r(mean) if missing(listprice)

* generate price_per_sqft
generate price_per_sqft = listprice / sqft

* coverting type in mumerical
encode type, gen (type_new)

 reg listprice baths_full beds sqft stories year_built price_per_sqft i.type_new

*generate log listprice
gen lnlistprice = ln(listprice)

*regression with lnlistprice
 reg lnlistprice baths_full beds sqft stories year_built price_per_sqft i.type_new // Adj R= 87.03

* Update city variable
replace city = regexs(1) if regexm(url, "_([A-Za-z\-]+)_([A-Z]{2})_(\d{5})_")
replace city = subinstr(city, "-", " ", .) // Replace dashes with spaces

* Update state variable
replace state = regexs(2) if regexm(url, "_([A-Za-z\-]+)_([A-Z]{2})_(\d{5})_")

* Update postal_code variable
replace postal_code = regexs(3) if regexm(url, "_([A-Za-z\-]+)_([A-Z]{2})_(\d{5})_")

* Verify the updated variables
list url city state postal_code if !missing(city)

* Coverting city into numerical
encode city, gen(city_num)

*regression with i.city_num and i.type_new
reg lnlistprice baths_full beds sqft stories year_built price_per_sqft i.type_new i.city_num // adj r= 95.97, r^2= 98.88

reg lnlistprice baths_full beds sqft stories year_built price_per_sqft i.type_new city_num // adj r= 87.68

reg lnlistprice i.city_num

// List all unique cities in the dataset
levelsof city, local(citylist)
display `"`citylist'"'

// Count the number of unique cities
```

```
egen unique_cities = group(city)
summ unique_cities

// Check the number of unique cities
distinct city

// List all unique cities and count them
levelsof city, local(citylist)
display "`: word count `citylist'"

// Calculate the percentage of missing values
count if missing(year_built)
di "Percentage missing: " r(N) / _N * 100

**NOV20**
****
clear all
cd "C:\Users\aindraba\Documents\final_project\datasets"
use "real_estate_final_data_NOV23.dta"

//linearity
estat ovtest

//trying
gen sqft_sq = sqft^2
reg lnlistprice baths_full beds sqft sqft_sq stories year_built price_per_sqft i.type_new i.city_num
gen beds_baths = beds * baths_full
reg lnlistprice baths_full beds sqft beds_baths stories year_built price_per_sqft i.type_new i.city_num


///omitted variable
estat ovtest
//multicolienarity
estat vif
//normality
predict res_std,rstandard
swilk res_std
////qq plot
qnorm res_std, rlopts(lcolor(red)) aspect(1)

// normality using k density (bell shaped curve)
predict res, resid
kdensity res, normal
qnorm res, rlopts(lcolor(red))
swilk res

reg lnlistprice baths_full beds sqft stories year_built price_per_sqft i.type_new i.city_num

///rerun for each variable/heteroscedacity
estat hettest
//Solution to heteroscedascity
reg lnlistprice baths_full beds sqft stories year_built price_per_sqft i.type_new i.city_num, vce(robust)

///outlier assumption
predict res_stud, rstudent


******************
*****23 Nov*******
******************
clear all
cd "C:\Users\aindraba\Documents\final_project\datasets"
use "real_estate_final_data_NOV23.dta"

// log variables
gen lnbaths_full = ln(baths_full)
gen ln_beds = ln(beds)
gen ln_sqft = ln(sqft)
gen ln_stories = ln(stories)
gen ln_yearbuilt = ln(year_built)
gen ln_pricesqft = ln(price_per_sqft)

reg lnlistprice lnbaths_full ln_beds ln_sqft ln_stories ln_yearbuilt ln_pricesqft i.type_new i.city_num // 1 adj r

**Trying out different variation of regression
```

```
reg lnlistprice lnbaths_full // significant adj r 37
reg  lnlistprice lnbaths_full ln_beds // ln_beds insignificant
reg  lnlistprice lnbaths_full ln_sqft // both significant adj r 55
reg lnlistprice lnbaths_full ln_sqft ln_stories // all Significant try ln_stories squared

reg lnlistprice lnbaths_full ln_sqft ln_stories ln_yearbuilt // adj r 54

reg lnlistprice lnbaths_full ln_beds ln_sqft ln_stories ln_yearbuilt //ln_stories not significant
reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt // all significant adj r 57
estat vif

reg lnlistprice lnbaths_full ln_beds ln_pricesqft // different model


reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt i.type_new i.city_num // adj r 72

reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt i.type_new // adj r 60
estat ovtest

** Generating interactive and polynomial terms
gen ln_sqft2 = ln_sqft^2
gen ln_yearbuilt2 = ln_yearbuilt^2
gen lnbaths_full2 = lnbaths_full^2
gen ln_beds2 = ln_beds^2

// Limited interaction terms
gen ln_sqft_baths = ln_sqft * lnbaths_full
gen ln_sqft_yearbuilt = ln_sqft * ln_yearbuilt
gen baths_yearbuilt = lnbaths_full * ln_yearbuilt
gen ln_sqft_type = ln_sqft * type_new
gen ln_sqft_lnbeds = ln_sqft * ln_beds
gen lnbaths_lnbeds = lnbaths_full * ln_beds


** Different regression Models
// Model 1
reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ///
    ln_sqft_baths ln_sqft_lnbeds ln_sqft_yearbuilt lnbaths_lnbeds i.type_new

//Model 2
reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ///
    ln_sqft2 ln_yearbuilt2 lnbaths_full2 ln_beds2 i.type_new

//Model 3
reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ///
    ln_sqft_baths ln_sqft_yearbuilt ln_sqft_lnbeds ///
    ln_sqft2 ln_yearbuilt2 lnbaths_full2 ln_beds2 i.type_new

//Model 4
reg lnlistprice lnbaths_full ln_beds ln_yearbuilt ///
    ln_sqft_type ln_yearbuilt2 ln_beds2 i.type_new

//Model 5
reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ///
    ln_sqft_baths ln_sqft_lnbeds ln_sqft_yearbuilt ///
    baths_yearbuilt lnbaths_lnbeds ///
    ln_sqft2 ln_yearbuilt2 lnbaths_full2 ln_beds2 i.type_new


**********
****BEST MODEL SO FAR***

//Model 4
reg lnlistprice lnbaths_full ln_beds ln_yearbuilt ///
    ln_sqft_type ln_yearbuilt2 ln_beds2 i.type_new

** Diagnostic checks
// normality using k densty (bell shaped curve)
predict res, resid
kdensity res, normal
// Shapiro–Wilk W test for normal data
predict res_std, rstandard
swilk res_std
```

```
//multicoliearnity
estat vif

//linearity
estat ovtest

//heteroscedacity
estat hettest

**Model 4 Robust**
reg lnlistprice lnbaths_full ln_beds ln_yearbuilt ///
    ln_sqft_type ln_yearbuilt2 ln_beds2 i.type_new, vce(robust)


reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ln_sqft_baths ln_sqft_lnbeds ln_sqft_yearbuilt lnbaths_lnbeds i.type_new
outreg2 using myreg.doc, replace ctitle (Model 1) label

reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ln_sqft2 ln_yearbuilt2 lnbaths_full2 ln_beds2 i.type_new
outreg2 using myreg.doc, append ctitle (Model 2) label

reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ln_sqft_baths ln_sqft_yearbuilt ln_sqft_lnbeds ln_sqft2 ln_yearbuilt2 lnbaths_full2 ln_beds2
i.type_new
outreg2 using myreg.doc, append ctitle (Model 3) label

reg lnlistprice lnbaths_full ln_beds ln_yearbuilt ln_sqft_type ln_yearbuilt2 ln_beds2 i.type_new, vce(robust)
outreg2 using myreg.doc, append ctitle (Model 4) label

reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ln_sqft_baths ln_sqft_lnbeds ln_sqft_yearbuilt baths_yearbuilt lnbaths_lnbeds ln_sqft2
ln_yearbuilt2 lnbaths_full2 ln_beds2 i.type_new
outreg2 using myreg.doc, append ctitle (Model 5) label

/////////
reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ln_sqft_baths ln_sqft_lnbeds ln_sqft_yearbuilt lnbaths_lnbeds i.type_new
outreg2 using myreg.doc, replace ctitle (Model 1) label dec(3) title (Table 1: Factors which effect the real estate price in Texas in 2024. )

reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ln_sqft2 ln_yearbuilt2 lnbaths_full2 ln_beds2 i.type_new
outreg2 using myreg.doc, append ctitle (Model 2) label

reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ln_sqft_baths ln_sqft_yearbuilt ln_sqft_lnbeds ln_sqft2 ln_yearbuilt2 lnbaths_full2 ln_beds2
i.type_new
outreg2 using myreg.doc, append ctitle (Model 3) label

reg lnlistprice lnbaths_full ln_beds ln_yearbuilt ln_sqft_type ln_yearbuilt2 ln_beds2 i.type_new, vce(robust)
outreg2 using myreg.doc, append ctitle (Model 4) label

reg lnlistprice lnbaths_full ln_beds ln_sqft ln_yearbuilt ln_sqft_baths ln_sqft_lnbeds ln_sqft_yearbuilt baths_yearbuilt lnbaths_lnbeds ln_sqft2
ln_yearbuilt2 lnbaths_full2 ln_beds2 i.type_new
outreg2 using myreg.doc, append ctitle (Model 5) label

label variable lnbaths_full "Log of Bathrooms"
label variable ln_beds "Log of Bedrooms"
label variable ln_sqft "Log of Square Feet"
label variable ln_yearbuilt "Log of Year Built"
label variable ln_sqft_baths "Log of Square Footage × Bathrooms"
label variable ln_sqft_lnbeds "Log of Square Footage × Bedrooms"
label variable ln_sqft_yearbuilt "Log of Square Feet × Year Built"
label variable lnbaths_lnbeds "Log of Bathrooms × Bedrooms"
label variable ln_sqft2 "Log of Square Feet^2"
label variable ln_yearbuilt2 "Log of Year Built^2"
label variable lnbaths_full2 "Log of Bathrooms^2"
label variable ln_beds2 "Log of Bedrooms^2"
label variable ln_sqft_type "Log Feet × Property Type^2"
label variable baths_yearbuilt "Bathrooms × Year Built"

label define type_new_lbl 2 "Farm" 4 "Mobile Home" 6 "Single Family" 7 "Townhomes"

label values type_new type_new_lbl

ssc install outreg2

* Summarize the variables of interest
summarize listprice baths_full beds sqft stories year_built type_new

outreg2 using summary_stats.doc, replace su m(log) word keep(listprice baths_full beds sqft stories year_built type_new)
```