Residential Real Estate: A Study of the Sale Price of Individually Owned vs. Non-individually Owned Condos

Master of Quantitative Economics (MQE)

University of California, Los Angeles (UCLA)

Author: George Musabandesu

Faculty Advisor: Dr. Randall R. Rojas

November 30, 2024

# Abstract

This study examines condominium sale prices in Cook County, Illinois, from 2014 to 2020, focusing on the impact of ownership type on property values. Specifically, it investigates whether condominiums, hereto referred as condos, owned by individuals sell at different prices compared to those owned by corporations, LLCs, or other non-individual entities. By leveraging several Cook County, Illinois datasets of condo sales, this research employs statistical analysis using linear regression and machine learning tools such as tree-based models like Random Forest, Decision Tree Regression, and XGBoost to assess price disparities based on ownership type, controlling for variables such as location, number of bedrooms, full baths, unit size, and condo construction year.

The analysis revealed that tree-based models yielded significant results and had a better performance compared to linear models. Tree-based models handled the complexity and intricacies of the imbalanced dataset and dealt effectively with the data's lack of a normal distribution and heteroskedasticity (non-constant variance of the residuals across all the predictor variables). The non-normality and heteroskedasticity conditions violated the assumptions required for linear regression analysis. Ordinary Least Squares (OLS) regression results indicated high levels of multicollinearity among the independent variables, necessitating the use of alternative regression models to mitigate this issue. Consequently, robust models such as Ridge regression and Lasso were deployed, yet their performance was not better than the initial OLS and they exhibited even higher Mean Square Errors (MSEs) and lower R-Squared results compared to non-parametric models. The tree-based models, which are well-suited for non-parametric data, produced lower MSEs than linear models. A significant test demonstrated a notable difference in condo sale prices between individual and non-individual owners, underscoring the impact of ownership type on property values. These findings contribute to a deeper understanding of how ownership affects real estate values.

**Table of Content**

[1. Abstract 1](#_Toc182773168)

[2. Introduction 3](#_Toc182773169)

[2.1 About Cook County, Illinois 4](#_Toc182773170)

[3. Literature Review 4](#_Toc182773171)

[4. Summary of existing research & theories 5](#_Toc182773172)

[5. Methodology 5](#_Toc182773173)

[6. Research Design & Approach 6](#_Toc182773174)

[7. Data collection/cleaning 7](#_Toc182773175)

[8. Data Analysis 8](#_Toc182773176)

[9. Results 9](#_Toc182773177)

[9.1 Tables/Charts/Graphs to showcase key points 9](#_Toc182773178)

[10. Discussion 17](#_Toc182773179)

[10.1 Interpretation of results 17](#_Toc182773180)

[10.2 Comparison with Existing Research 17](#_Toc182773181)

[10.3 Implications of the Findings 17](#_Toc182773182)

[10.4 Limitations of the Study 18](#_Toc182773183)

[11. Conclusion 18](#_Toc182773184)

[12. References 19](#_Toc182773185)

# Introduction

Condominiums, or condos, are individually owned units within a larger residential complex. Owners have rights to their unit's interior but share common areas, such as hallways and exterior spaces, with other residents. A management group or a board of unit owners typically oversees the complex’s operations. Condo ownership comes with responsibilities, including monthly HOA fees for property maintenance and potential special assessments for major repairs. Ownership is the key difference concerning condos and apartments. Condos are owned by individuals, while apartments are rental units owned by a single entity, such as a landlord or property management firm. Condo owners pay property taxes and HOA fees, whereas apartment renters pay rent while the building owner covers taxes and property fees. Real estate markets are influenced by a variety of factors, from economic conditions to property characteristics and ownership types. The real estate market is a multifaceted realm that reflects the multipart interplay of economic, social, and policy-driven factors. Understanding the determinants of property prices is essential for stakeholders ranging from policymakers and investors to potential homeowners. A particular area of interest in this field is the influence of ownership type on the sale price of properties. Several news articles including CNBC (Sola 2024), New York Post (Realtor.com, 2024), and a publication by the Joint Center for Housing Studies of Harvard University (JCHS) have discussed how there has been an increase in residential property investments by non-individual owners and some implications for affordability for individual owners. This trend raises questions about how ownership type might impact property values and affordability. Are condos owned by individuals valued differently than those owned by corporations, LLCs, or other non-individual entities? Answering this question is vital for stakeholders, including investors, policymakers, and prospective homeowners, as it sheds light on potential biases in the real estate market and can guide investment and regulatory choices. Previous studies on real estate pricing have often focused on macroeconomic influences, neighborhood characteristics, or individual property attributes. However, the impact of ownership type on sale price remains less explored, particularly for condominium sales. By examining this question, this research aims to contribute to the broader understanding of how different market actors may influence pricing dynamics in the housing sector.

This research paper investigates this phenomenon within the condo market of Cook County, Illinois, over a seven-year period from 2014 to 2020. By analyzing a comprehensive dataset of condo sales, this study aims to identify any significant price differentials associated with ownership types while controlling for factors such as unit size, age, and location. The results of this analysis will contribute to the literature on property ownership and real estate valuation, offering insights that could inform future investment strategies and policy discussions around corporate versus individual condominium property ownership.

Leveraging a range of statistical and machine learning techniques, including Ordinary Least Squares (OLS) regression, Ridge, Lasso, and advanced tree-based models such as Random Forest and XGBoost, this analysis seeks to provide a robust assessment of the relationship between condo ownership type and sale price. This comprehensive approach not only illuminates potential price discrepancies but also offers insights into the predictive capabilities of various models for informing future market analyses and decision-making.

## About Cook County, Illinois

Located in northeastern Illinois, Cook County is an urban area comprising over 800 local government units. With a population of approximately 5.2 million (2015 census), it is the second most populous U.S. County and ranks as the 19th largest government entity (<https://www.cookcountyil.gov/about-cook-county>).

The county was established on January 15, 1831, by the Illinois State Legislature, and became the 54th county in the state. The City of Chicago is the county seat and administrative hub. According to the county website, “Chicago and the surrounding suburbs account for 85% of the county's 946 square miles”.

# Literature Review

There have been several research articles directed at the housing market with a focus on using robust statistical methods to analyze and predict sale prices. Ordinary Least Squares (OLS) regression was used to analyze the factors of housing prices, with a specific focus on the effects of regulations in the land and housing market (Malpezzi 1996). The method used involved models of house prices and rents in what their research described as a ‘simple supply-and-demand framework focusing on incomes, population changes…” Other statistical models capable of dealing with non-linear and complex relationships include Lasso and Ridge regressions. Least Absolute Shrinkage and Selection Operator (lasso), is a type of linear regression that includes an L1 penalty term. This penalty term has the effect of shrinking some of the regression coefficients towards zero, which can help with feature selection (by making some coefficients equal to zero) where many features are present, and regularization, where the L1 penalty terms assists with reducing overfitting through addition of a constrain to the model. Lasso regression was the subject of an article that focused on discussing a new method for estimation in linear models (Tibshirani 1996). The lasso regression was used in a research paper that aimed to predict house prices with more efficiency and accuracy (Sharma et. al.. 2024). This paper used both Lasso and Ridge regressions in the analysis of housing prices and found these regression techniques useful as there are several factors that would impact housing prices. Several articles have also discussed the use of Tree-based models in predicting housing prices. The popularity in the use of machine learning in predicting house prices keeps on rising. A Boston housing dataset was analyzed using Random Forest regression to predict the prices of houses based on the features in that dataset (Adetunji et al, 2022). Random Forest’s capability of combining multiple decision trees in conducting predictions makes it very desirable as its outcomes are more stable than just using once decision tree.

# Summary of existing research & theories

Several statistical techniques have been used in predicting house prices. From Linear regressions, such as OLS to regularized regressions like Lasso and Ridge regressions, to Tree-based models like Random Forest and Decision Tree regressions. Most of the existing research and theories about the housing market have tended to lean heavily on price prediction for specific types of housing, using specific well-known features like house size, location, number of bedrooms, year built and the like. The effectiveness and efficiency of the use of these statistical techniques are well known. When it comes to using these techniques to look at ownership type and how this might influence the sale price, especially in the condominium market, the research seems lacking. One thesis study focused on asking the question of who was buying the condominiums in the four regions that included Atlanta, Boston, Chicago, and San Diego (DeLaney, Pizzuti, 2005) and did not look at the composition of the sellers of these units and how they arrived at the price points they offered.

# Methodology

The research methodology incorporated rigorous data pre-processing to create features identifying seller ownership types, followed by exploratory data analysis (EDA) to uncover necessary data patterns. The study aimed to determine if condo sale prices varied based on ownership type. The dataset from Cook County, Illinois, was meticulously cleaned to remove irrelevant variables and handle null values. EDA involved the use of correlation plots, pair plots, histograms, and box plots to examine feature interactions. A time series plot depicted the mean and median condo prices based on ownership type from 2014 to 2022, revealing market trends. Further statistical analysis included the Jarque-Bera test for normality, the Breusch-Pagan test for heteroscedasticity, and t-tests to compare mean and median condo sale prices by ownership type.

The study began with Ordinary Least Squares (OLS) regression to establish a baseline model, followed by Ridge and Lasso regressions to address potential multicollinearity among independent variables. Tree-based models, such as Random Forest, Decision Tree, and XGBoost, were employed to capture complex feature interactions beyond linear models. These robust techniques provided a comprehensive understanding of how ownership type influences condo sale prices**.**

# Research Design & Approach

The goal of the research is an analysis of the differences in condo sale prices in Cook County, Illinois based on ownership type. We want to understand if ownership type – Individual or Non-Individual (e.g., corporate, or institutional seller) has a significant influence on the sale price, considering all the other factors that could influence the sale price. Potential influential factors include the size of the condo, its location, when it was built, how many bedrooms and bathrooms, and ownership type. The analysis uses a comprehensive dataset of condo sale prices from 2014 to 2020. This timeframe and choice of features to analyze ensures that the model accounts for these property-related settings and regional variations. The research utilizes various statistical techniques using linear regression methods such as Ordinary Lease Squares (OLS), Ridge, and Lasso regression. The investigation also applies machine learning techniques such as Random Forest Regression, Decision Tree Regression and XGBoost to enhance the robustness and reliability of the findings. OLS regression is used as a means of baseline understanding of the coefficient relationships with the target variable, sale price. To deal with the imbalanced data, bootstrap is employed in OLS to see if resampling of the dataset produces significant changes in results before employing other additional regression techniques. Regularized models such as Ridge and Lasso are implemented to address multicollinearity and enhance predictive accuracy. Additionally, tree-based models such as Random Forest, Decision Tree, and XGBoost are utilized in the capture of non-linear interactions and complex variable relationships.

A systematic approach is used to evaluate each model’s performance. Diagnostic tests and residual plots are used in verification of key assumptions of linear regressions such as normality, homoscedasticity, and independence of errors. Additionally, these evaluations are further complemented by analysis and evaluation of the predictive power of the models by capturing metrics including Mean Square Errors (MSE) and R-squared values. Tree-based models also provide results on feature importance and how each feature contributes to predicting the sale price. We also conduct a t-test based on the results of Random Forest Regression to establish whether there is a statistical difference in condo sale price based on ownership type. The use of all these statistical techniques ensures a thorough understanding of how ownership type influences condo prices and how reliable these findings are across different models. These approaches provide a robust analysis capable of informing stakeholders about pricing demand and market trends centered on the Cook County condo market.

# Data collection/cleaning

The two datasets used in the research came from the Cook County website. One dataset is called the Assessor – Parcel Sales (<https://datacatalog.cookcountyil.gov/Property-Taxation/Assessor-Parcel-Sales/wvhk-k5uv/about_data>) which contains data for the parcel sales from 1999 to about 2020. This dataset was created in 2022, and the last update wan in October 2024. This dataset contains about 2.5 million rows with each row representing a sale, and 19 columns. The second dataset used in this research is called The Assessor – Residential Condominium Unit Characteristics (<https://datacatalog.cookcountyil.gov/Property-Taxation/Assessor-Residential-Condominium-Unit-Characterist/3r7i-mrz4/about_data>). This dataset was created in April 2022 with the last update on November 15, 2024. It contains about 10.5 million rows with each row representing a condominium Unit. There is a total of 24 columns. These two datasets all have the column called PIN, Parcel Identification Number – which was used to link them into one dataset. The first step of data cleaning was removing all the null values in the dataset especially in the columns that identified the seller. Several columns were removed from the dataset as they did not contribute meaningfully to the research. Also, other removed rows included transactions that where repeated and several transactions with missing sale price or those with a sale price of zero. Other transactions had a sale price of $1 and all the transactions with a sale price below $10,000.00 were also excluded from the dataset. Due to the high volume of the transactions the research period was limited to cover sale prices from 2014 to 2020. Some data cleaning techniques included identification of seller type, whether individual or non-individual and creating a categorical variable with a binary outcome of 1 if seller is individual and 0 otherwise. Once the data cleaning process was finalized the remaining dataset had a total of 25,834 rows for the rest of the analysis.

# Data Analysis

This research’s data analysis utilized a comprehensive exploration of the Cook County, Illinois condo sales dataset which was a combination of two different datasets from the same source. The focus of the analysis was on the potential differences in condo sale prices based on ownership type. First, we conducted a preliminary examination of the dataset’s structure followed by creation of visual and statistical insights into the data characteristics. Feature relationships were explored using correlation plots, pair plots, box plots, and histograms. The sale price was further analyzed through individual histograms and a data transformation was conducted to simulate a more normal distribution. In observing outliers, the box plots were utilized to show the log sale price based on ownership type. A time series plot helped in highlighting the market trend of the mean and median sale price based on ownership type. Histograms were also useful in showing segmented representations of counts of sales based on ownership type. The discovery of the existing imbalance in the dataset prompted the use of a bootstrapping technique to balance the dataset. An OLS regression utilizing bootstrap techniques did not yield much significant difference from the OLS without bootstrap techniques, leading to the conclusion that further analyses using regularized methods were needed. The results of the OLS regression revealed high levels of multicollinearity of the predictors and so a Variance Inflation Factor (VIF) metric test was conducted to identify and potentially exclude features that showed high multicollinearity. Several township location codes recorded high VIF greater than 5, indicating high correlation between those predictors in relation to other predictors. Diagnostic tests such as Jarque-Bera and Breusch-Pagan tests revealed the non-normality and heteroskedastic nature of the dataset, discovering that pointed to further concerns about the significance of the regression results. Further tests looked for any influential, high leverage, and outlier data points that had the potential to skew the results.

Beyond the regularized regressions, tree-based models like Random Forest, Decision Tree, and XGBoost were used and their efficiency of working with complex data lead to lower MSEs and Higher R-squared scores compared to linear models. These tests further showed how certain features like unit square footage and location, followed by ownership type were critical to the sale price predictions based on ownership type. In relation to the results from the VIF metrics, the township codes that random forest ranked as important also had VIF readings that were below 5, which is acceptable in terms of correlation and multicollinearity.

T-tests were conducted based on the results of the overall best fitting model, to show that there was a statistically significant difference in condo sale price based on ownership type, with individual sellers showing a higher price over non-individual sellers.

# Results

## Tables/Charts/Graphs to showcase key points

A graph of a sales chart

Description automatically generated with medium confidence

Figure Boxplot of Log Sale Price by Seller type

This boxplot shows a comparison of log sale price between individual sellers and non-individual sellers. 1 corresponds to individual sellers and 0 to non-individual sellers. The distribution of the two plots shows a somewhat similar distribution except for the presence of outliers. The interquartile range (IQR) is higher for individual sellers compared to non-individual sellers. The non-individual sellers show several outliers.

A graph of a distribution of sales

Description automatically generated

Figure Frequency of Log Sale Price by Seller type

The above histogram visualizes the distribution of Log sale price based on ownership type. Individual sellers show a bimodal distribution and the non-individual sellers’ distribution shows skewness and the presence of outliers.

A graph of different colored squares

Description automatically generated

Figure Boxplots of Log Sale Price by number of bedrooms by Seller type

This box plot shows the distribution of log sale prices of condos categorized by the number of bedrooms and distinguished by seller type (individual sellers vs. non-individual sellers). Price Distribution by Bedroom Count: The overall log sale price generally increases with the number of bedrooms, which is expected as larger units typically command higher prices. The spread of sale prices (indicated by the interquartile range) is wider for units with 3 or more bedrooms, suggesting greater variability in the sale prices of larger condos. The comparison between seller types shows that for smaller units (0–2 bedrooms), median sale prices are similar for individual (orange) and non-individual sellers (blue), with individual sellers sometimes achieving slightly higher prices. For larger units (3+ bedrooms), individual sellers tend to have higher median log sale prices. Price variability is more pronounced for non-individual sellers in smaller units, suggesting greater price fluctuations. For units with 5+ bedrooms, individual sellers typically have higher prices, although data is sparse. These findings imply that individual sellers may secure higher prices, especially for larger condos, and non-individual sellers face more price variability, particularly for smaller units. Further statistical analysis confirmed the significance of these trends.

A graph of sales and performance

Description automatically generated with medium confidence

Figure Normalize sp500 based on log condo sale price

The normalized S&P 500 in the above graph shows consistent growth from 2014 to 2021, with significant volatility during the 2020 COVID-19 crash and recovery. In contrast, condo sale prices remain relatively stable with occasional spikes, particularly for non-individual sellers (type 0). There ais little correlation between the S&P 500's performance and condo prices, indicating that condo prices in this dataset are influenced by factors other than broad market trends. Non-individual sellers show some spikes of higher prices than individual sellers in some years, but individual sellers show higher prices overall.

A screenshot of a graph

Description automatically generated

Figure Correlation Plot

The correlation heatmap provides key insights into the factors influencing sale prices. Notably, there are strong positive correlations between log\_sale\_price and unit\_sf (0.61) as well as num\_bedrooms (0.47), indicating that larger units and more bedrooms generally command higher prices. Additionally, a moderate correlation is observed between log\_sale\_price and num\_full\_baths (0.49), suggesting that an increase in the number of bathrooms also tends to raise sale prices. In contrast, weak correlations are found with year\_built (0.006) and Is\_Individual\_Seller (0.24), showing that these factors have minimal impact on sale prices. Inter-feature correlations, such as those between num\_bedrooms and unit\_sf (0.69) and num\_full\_baths and unit\_sf (0.63), indicate that larger units tend to have more bedrooms and bathrooms. Overall, square footage and the number of bedrooms are strongly linked to higher sale prices, while factors like the year built and ownership type show weaker correlations. Additionally, it is noted that some locations yield positive correlations with sale prices while others do not, which is a common occurrence in the real estate market.

A screenshot of a graph

Description automatically generated

Figure Pairplots

The above pairplot displays the distribution of individual features (diagonal) and pairwise relationships between features (off-diagonal). Positive correlations (e.g., unit\_sf and log\_sale\_price) show an upward trend, while weak or no correlations (e.g., num\_half\_baths and year\_built) lack a clear pattern. Clustering patterns may suggest subgroups within the data, while outliers indicate unusual data points. Key feature insights include the positive correlation between log\_sale\_price and unit\_sf, weaker correlations with year\_built, and potential clustering of individual vs. non-individual sellers. The pairplot helps identify correlations, patterns, and outliers, guiding further analysis for feature selection and model building.

A graph with text on it

Description automatically generated

Figure Random Forest Feature Importance

**Model Performance:**

|  |  |
| --- | --- |
| Cross-validation MSE scores: | [0.1770565 ,0.18110168, 0.17309757, 0.16504415 0.17347227] |
| Mean Cross-validation MSE | 0.1740 |
| Mean Squared Error (MSE) | 0.1616 |

Figure : Table of results from Random Forest Regression Analysis

This cross validation test indicates that the average squared difference between the predicted and actual sale prices is relatively low, suggesting a good fit. The model has a good fit as its MSE (0.1616) is close to the cross validation MSE (0.1740).

**Feature Importance:** The importance graph highlights the relative impact of each feature on the model's predictions. The most influential feature is unit\_sf, with a score of 0.605421, indicating that larger units generally command higher prices. The year\_built, with a score of 0.224564, also plays a significant role, suggesting that newer units may be valued more highly. Various township codes, such as township\_code\_12 (0.011902), township\_code\_32 (0.015020), and township\_code\_37 (0.039809), reflect the impact of location on sale prices. Although less influential than square footage or location, the Is\_Individual\_Seller feature, with a score of 0.037911, still has a noticeable impact on the sale price. This analysis underscores the importance of unit size and construction year, while also highlighting the varying influence of location and ownership type on property values.

**Other Features:** The key features influencing condo sale price include unit size (unit\_sf), year built, ownership type (Is\_Individual\_seller), and location (township\_code), followed by the number of bedrooms and baths. The model demonstrates strong generalization, as the cross-validation MSE is consistent with the training/test MSE, indicating reliable performance on unseen data. Cross-validation confirms the model’s stability and suggests it will perform similarly on future data with an MSE around 0.1740. The Random Forest model, with an MSE of 0.1616 and R-squared of 0.82, outperforms the OLS and Ridge models, explaining over 82% of the variability in sale price, indicating a more accurate prediction.

A screen shot of a graph

Description automatically generated

Figure SHAP results showing impact of features on model prediction

The SHAP plot above illustrates the impact of various features on the model's predictions, with each point representing a SHAP value for a single prediction. The SHAP plot provides valuable insights into how each feature influences the model's predictions. The color gradient helps identify the feature values that have the most significant impact, with red points typically contributing positively and blue points contributing negatively to the predicted sale prices.

**Feature Importance:** The features are ranked by their importance, with the most impactful features at the top. unit\_sf, Is\_Individual\_Seller, and num\_full\_baths appear to be the most important features influencing the model's predictions.

A diagram of a diagram

Description automatically generated

Figure Random Forest Decision Tree Diagram

In the decision tree, seller type (represented by the Is\_Individual\_Seller feature) affects sale price predictions at specific splits:

The feature *Is\_Individual\_Seller* appears as a middle-level node in the decision tree. Specifically, it splits under the condition *township\_code\_37 <= 0.5*, where *Is\_Individual\_Seller <= 0.5* (non-individual sellers) leads to different subsequent nodes than when it is greater than 0.5 (individual sellers). This split impacts predictions by directing samples into subgroups with distinct predicted sale prices. The presence of *Is\_Individual\_Seller* in the path shows its role in influencing the log sale price, alongside features like *year\_built*, township codes, and the number of bedrooms.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Model** | **MSE** | **R-squared** |
| 1 | OLS (bootstrap) | 0.3097 | 0.6530 |
| 2 | Ridge | 0.4496 | 0.4965 |
| 3 | Lasso | 0.5416 | 0.3934 |
| 4 | Decision Tree | 0.211335 | 0.761832 |
| 5 | Random Forest | 0.161600 | 0.817900 |
| 6 | XGBoost | 0.184486 | 0.792090 |

Figure Summary of Outputs for all models used in the analysis

# Discussion

## Interpretation of results

The analysis revealed that there is a notable difference in condo sale prices based on ownership type. Specifically, properties sold by individual owners were found to have significantly different sale prices compared to those sold by non-individual entities. The model comparison showed that Random Forest yielded the best performance among all the models tested, achieving a Mean Squared Error (MSE) of 0.1616 and an R-squared value of 0.82. This high R-squared indicates that the model explained 82% of the variance in the condo sale prices, suggesting that the chosen features, including ownership type, were highly predictive of price. The final t-test conducted using the Random Forest model confirmed that the observed difference in sale price based on ownership type was statistically significant, reinforcing the validity of this finding.

## Comparison with Existing Research

Research on real estate pricing has largely focused on macroeconomic variables, property characteristics, and neighborhood factors. However, there is limited literature examining the specific impact of condo ownership type on sale prices. This study contributes to closing that gap by providing empirical evidence that condo ownership type does indeed play a role in determining the sale price of condos. While previous studies may have hinted at ownership dynamics influencing market behavior in general, few have quantified this effect in the context of condos, particularly within the Cook County, Illinois, region. Therefore, this research stands out by addressing an underexplored facet of real estate economics and adds nuance to the understanding of price determinants based on ownership type.

## Implications of the Findings

The implications of these findings are significant for various stakeholders. For potential buyers, understanding that ownership type can influence sale price may guide their negotiation strategies and decision-making processes. For investors and real estate professionals, recognizing the impact of ownership type on pricing could inform investment decisions, risk assessments, and portfolio management. Policymakers may also find this research useful in designing regulations or programs that ensure fair market practices, considering how ownership structures may affect housing affordability and price stability. These implications highlight the underexplored complexity that exists in the condominium real estate market based on ownership type.

## Limitations of the Study

Despite the compelling results, this study has some minor limitations. The dataset was imbalanced with individual sellers taking the majority stake. The tree-based models’ performance could be further enhanced by incorporating class weight adjustments that allows for giving more importance on the minority class to deal with this imbalance. Another approach would be a resampling technique using Synthetic Minority Over-Sampling Technique (SMOTE) that creates synthetic samples to improve the majority class. Moreover, the dataset was confined to Cook County, Illinois, which may limit the generalizability of the findings to other regions with different market dynamics. The study also focused solely on condos, so the results may not extend to other types of properties. Future research could expand on these findings by incorporating additional data from multiple geographic locations and by employing more complex machine learning techniques with higher computational capacity to validate and strengthen the conclusions.

# Conclusion

The analysis conducted on the Cook County, Illinois condo sale data from 2014 to 2020 provided valuable insights into the influence of ownership type on sale prices. The main findings indicate that there is a significant difference in condo sale prices based on whether the seller is an individual or a non-individual entity. Among the various predictive models applied, the Random Forest model yielded the best performance, achieving a Mean Squared Error (MSE) of 0.1616 and an R-squared value of 0.82. A subsequent t-test validated these results, confirming a statistically significant difference in condo sale prices by ownership type. While this study contributes to the understanding of market dynamics within the condominium sector, it also highlights the need for further research that incorporates wider geographic reach. Future studies could incorporate more advanced ensemble and bootstrap techniques on tree-based models to improve model robustness, Additionally, expanding the analysis to include other property types or broader geographical regions could provide more comprehensive insights. Overall, these findings underscore the importance of recognizing ownership type as a factor that may affect pricing in the real estate market. This has potential implications for investors, policymakers, and real estate professionals who seek to make informed decisions. The results of this research encourage a more nuanced view of ownership type and its influence on property valuations. Further exploration into the causal mechanisms behind these price differences would enrich the literature and guide more effective market strategies. Overall, this study conclusively provides significant evidence that there is a difference in condo sale prices based on ownership type.

# References

Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, Gbenle Oluwadara, House Price Prediction using Random Forest Machine Learning Technique,Procedia Computer Science, Volume 199, 2022

Cook County, Illinois Government. (n.d.). *Assessor Parcel Sales*. Open Data(<https://datacatalog.cookcountyil.gov/Property-Taxation/Assessor-Parcel-Sales/wvhk-k5uv/about_data>

Cook County, Illinois Government. (n.d.). *Assessor Residential Condominium Unit Characteristics.* Open Data. <https://datacatalog.cookcountyil.gov/Property-Taxation/Assessor-Residential-Condominium-Unit-Characterist/3r7i-mrz4/about_data>

DeLaney, Robert M. (Robert Michael); Pizzuti, Linda *Who is buying urban condominiums? a tale of four cities* K 2005 https://dspace.mit.edu/handle/1721.1/33200

Hermann, A. (2023, July 18). *8 facts about investor activity in the single-family rental market*. Joint Center for Housing Studies of Harvard University. <https://www.jchs.harvard.edu/blog/8-facts-about-investor-activity-single-family-rental-market>

Malpezzi, S. (1996). Housing Prices, Externalities, and Regulation in U.S. Metropolitan Areas. *Journal of Housing Research*, *7*(2), 209–241. <http://www.jstor.org/stable/24832860>

M. Sharma, R. Chauhan, S. Devliyal and K. R. Chythanya, "House Price Prediction Using Linear and Lasso Regression," *2024 3rd International Conference for Innovation in Technology (INOCON)*, Bangalore, India, 2024, pp. 1-5, doi: 10.1109/INOCON60754.2024.10511592.

Robert Tibshirani, Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, Volume 58, Issue 1, January 1996, Pages 267–288, <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>

Realtor.com. (2024, July 11). *Real estate investors are snapping up a record number of homes—especially in these 5 affordable cities*. New York Post. <https://nypost.com/2024/07/11/real-estate/real-estate-investors-are-snapping-up-a-record-number-of-homesespecially-in-these-5-affordable-cities/>

Solá, A. T. (2024, June 4). *Investor home purchases jump for the first time in two years. Here’s what that means for buyers*. CNBC. <https://www.cnbc.com/2024/06/04/investor-home-purchases-jump-for-the-first-time-in-two-years.html>