Cook County Housing Project Report

Abstract

This project investigates housing sale data from Cook County with a dual focus: building predictive models for property valuation and examining the fairness of those models in real-world contexts. Through exploratory analysis, feature evaluation, and visualization, I assessed which housing attributes most strongly influence sale prices. I also critically analyzed systemic issues in property tax assessments, including historical inequities and racial disparities. By comparing model performance across price ranges and interpreting residual plots, I found evidence of regressive valuation patterns, where lower-priced homes are systematically overestimated. This project highlights that while statistical accuracy (e.g., low RMSE) is important, true fairness requires attention to bias, transparency, and equity in both model design and the surrounding institutional processes.

Part 1: Data Understanding and Exploration

I began by identifying the granularity of the dataset. Each row represents a single property sale, with details such as lot size, number of bedrooms, wall/roof materials, neighborhood code, and sale history.

I speculated that this data was collected to track property characteristics and market trends. This information could be valuable to city planners, real estate professionals, and potential homebuyers.

To explore the dataset, I posed and investigated a few questions:

1. Which month sees the most property sales?

   I found that August was the most common sale month by calculating the mode of the 'Sale Month of Year' column.

2. What's the distribution of property age (by decade)?

   I discussed calculating the interquartile range using a histogram on the 'Age Decade' column.

3. Do older people tend to buy older homes?

   If I had access to demographic data, I would explore this using a scatter plot and KDE plot to examine the relationship between buyer age and property age.

I also noticed that the sale price distribution was heavily skewed, likely due to outliers such as luxury properties. I proposed applying a log transformation or selectively removing outliers to improve data representation.

To evaluate potential predictive features, I:

1. Created a jointplot of Log Building Square Feet vs. Log Sale Price (Figure 1) and observed a strong, linear relationship, indicating that Log Building Square Feet could be a valuable predictor.
2. Built a boxplot of Log Sale Price by number of bedrooms (Figure 2) to avoid overplotting and visualize how the number of bedrooms might relate to housing value.

Figure 1. Jointplot of Log Building Square Feet vs. Log Sale Price

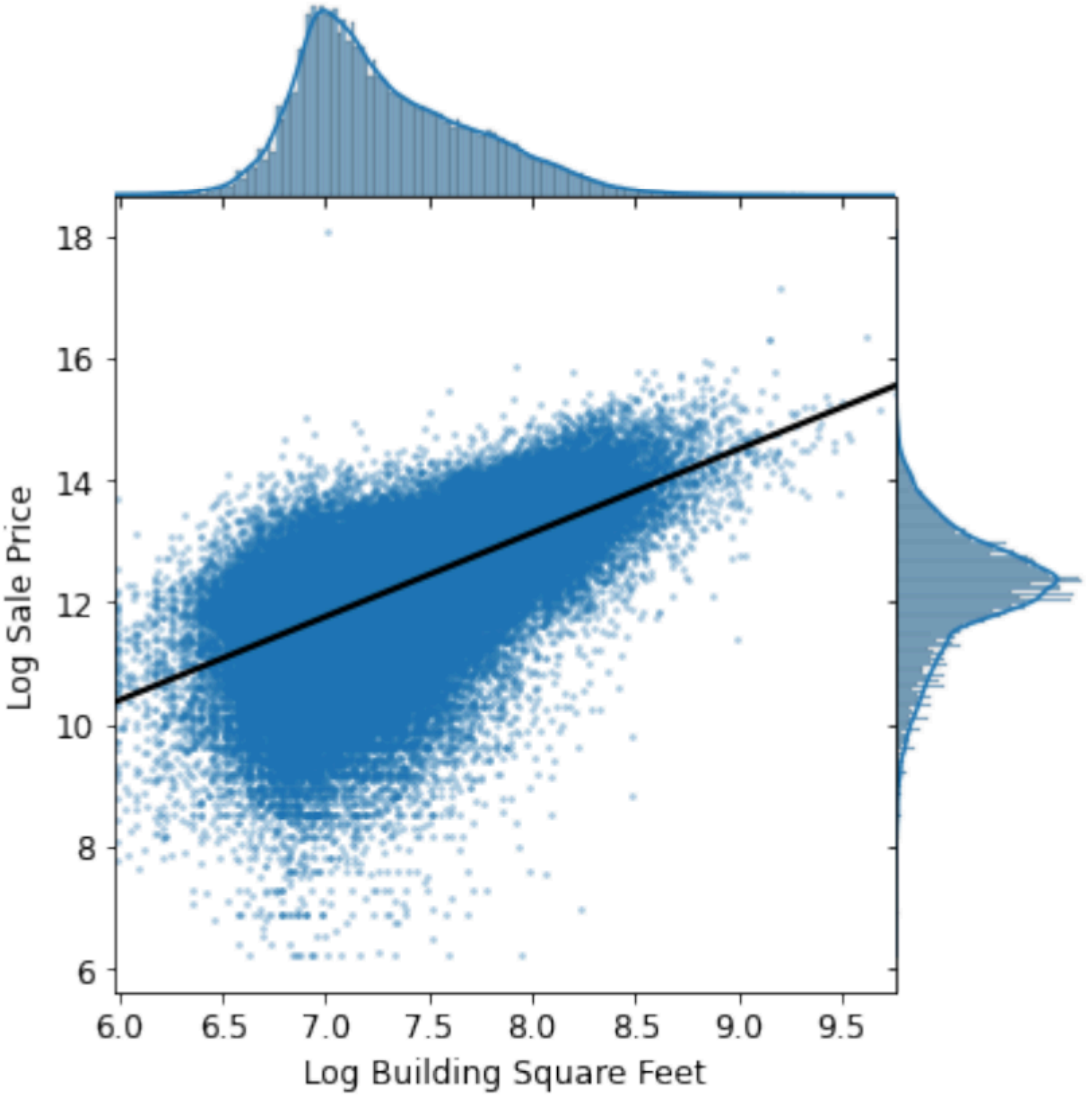
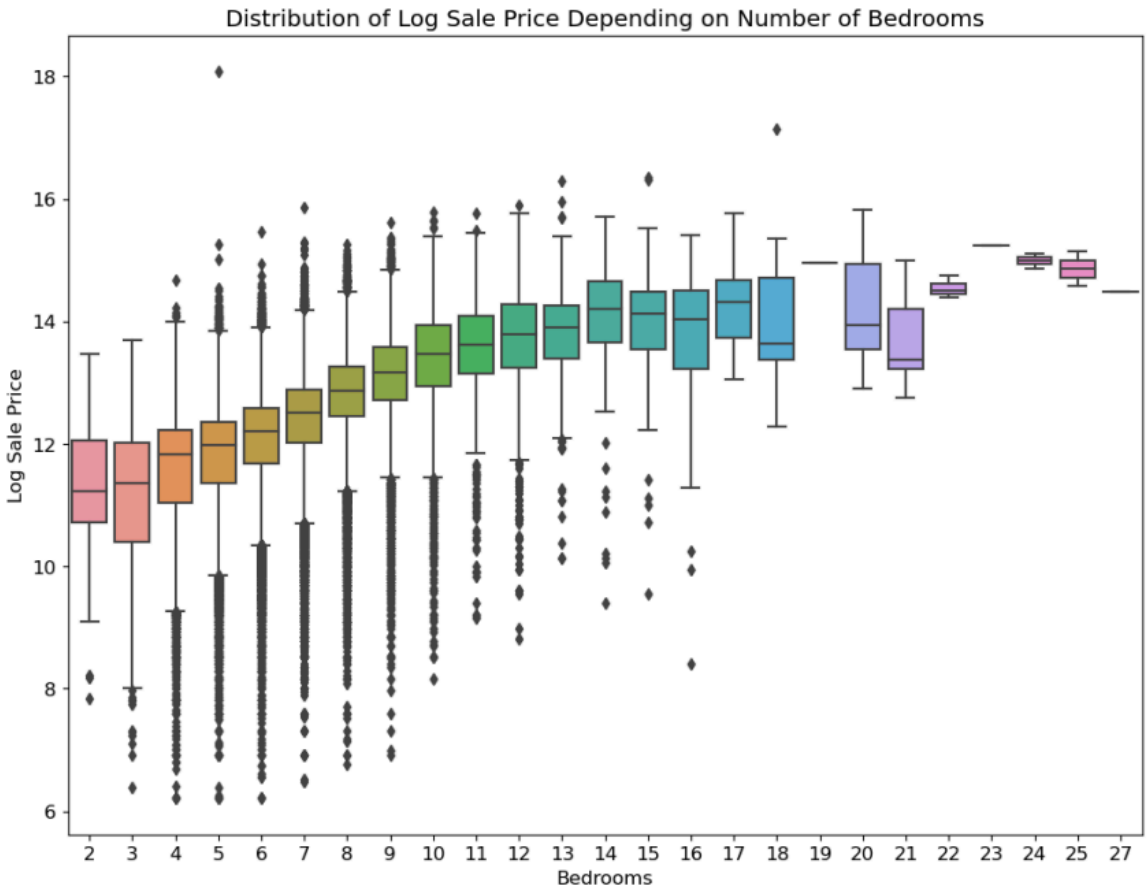


Figure 2. Boxplot of Log Sale Price by Number of Bedrooms



Part 2: Human Context, Ethics, and Fairness

In the second part of the project, I focused on the human context of housing valuation by exploring fairness and who is impacted by pricing decisions.

I considered stakeholders who care about the question “How much is a house worth?”:

1. Buyers, who are looking for affordability
2. Sellers, who want to maximize returns
3. Real estate agents, whose commissions depend on higher sale prices

I reflected on fairness in property assessments and found Scenario C (where inexpensive homes are overvalued and expensive homes are undervalued) to be the most unfair. This disproportionately affects low-income buyers, who are more likely to be priced out of homes they otherwise could afford.

Drawing from the Chicago Tribune's investigation into Cook County's tax system, I summarized key issues:

1. Inaccurate and biased assessments that failed fairness standards
2. Higher tax burdens on low-income and non-white homeowners
3. Unequal access to the appeals process, which favored wealthier and whiter neighborhoods

From a modeling perspective, I analyzed residuals vs. original Log Sale Price (Figure 3) and examined how RMSE (Figure 4) and the proportion of overestimated homes changed across price intervals (Figure 5).

Figure 3. Residuals vs. Original Log Sale Price

Figure 4. RMSE by Log Sale Price Interval

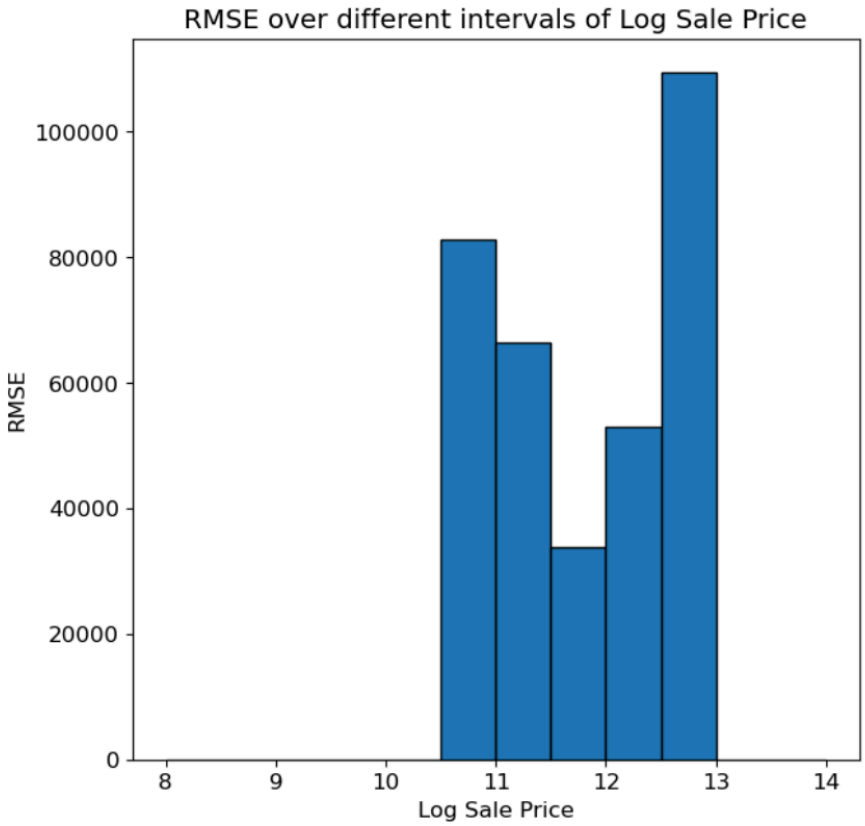
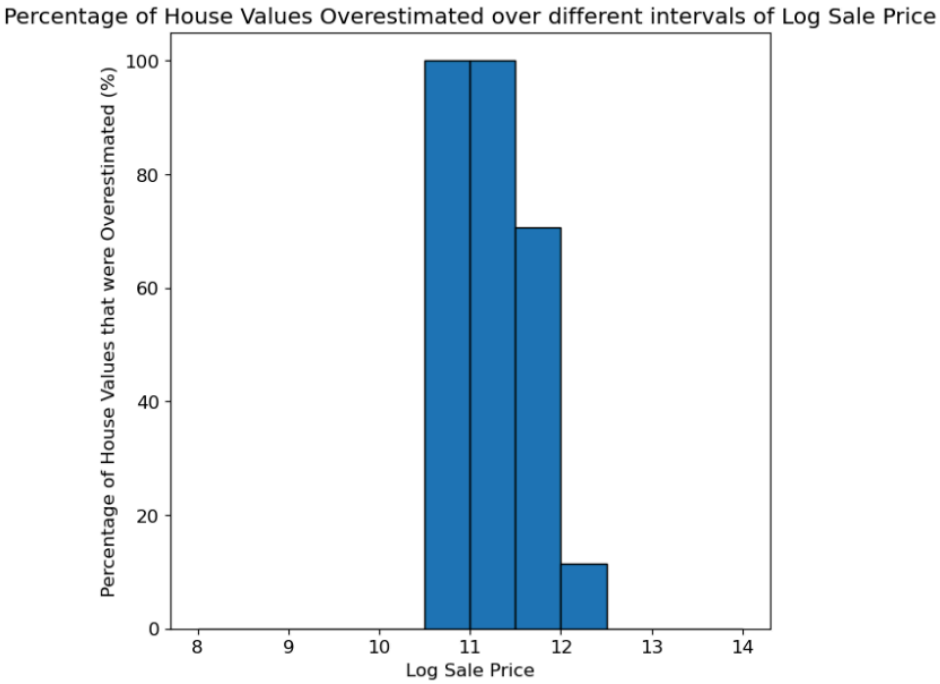


Figure 5. Percentage of Overestimated Properties by Log Sale Price Interval



Final Thoughts on Fairness and Accuracy

While a low RMSE indicates good overall predictive accuracy, it doesn't guarantee fairness. A fair model should:

1. Provide consistent valuations for similar properties
2. Be transparent and interpretable
3. Avoid systematic bias, particularly against low-income or marginalized groups

Fairness isn't just about numbers. It also depends on the surrounding systems. If appeals are only accessible to certain communities, even a statistically accurate model can result in unjust outcomes.

Overall Takeaway on Fairness in Cook County Housing

The Cook County property tax assessment system was systematically unfair, especially to low-income and non-white homeowners.

Key findings:

1. Biased assessments: Lower-income neighborhoods were often overvalued, while higher-income areas were undervalued—leading to higher tax burdens for those least able to afford them.
2. Unequal appeals: Wealthier, whiter homeowners were more likely to appeal and win lower assessments. Others lacked the time, money, or access to navigate the process.
3. Systemic inequity: These patterns reinforced historical segregation and economic inequality, treating homes "equally" in the model while ignoring unequal lived realities.
4. Accuracy ≠ Fairness: Even models with strong statistical performance can still reinforce inequity if their errors fall disproportionately on disadvantaged groups.

In short: Without thoughtful design and oversight, data-driven systems like property assessments can unintentionally amplify racial and economic disparities rather than correct them.