

# North Chicago Township Home Sale Analysis

## I. Executive Summary

In the competitive real estate market of North Chicago, where location is traditionally considered the prime determinant of property value, this project aims to delve deeper into the nuanced factors influencing the sale prices of single-family homes. The North Chicago Township, nestled within the City of Chicago and featuring affluent neighborhoods like Gold Coast, Magnificent Miles, and Lincoln Park, provides a rich landscape for our analysis. The study will employ a linear regression model on the Sale Price of these homes considering various continuous predictors.

By exploring the interplay of these factors, we aim to uncover valuable insights that go beyond the conventional wisdom of "location, location, location." This analysis will empower prospective homebuyers, sellers, and real estate professionals with a more comprehensive understanding of the intricate dynamics influencing property prices in the North Chicago Township.

## II. Data Description

Data Definition	Data Source	Data Size	Rows/Columns
North Chicago Township Home Sale data	<a href="#">Cook County Assessor 2021</a>	82 KB	(404, 28)

This data consists of 403 single-family homes sold in the North Chicago township of Cook County between 2018 and 2020.

## III. Business Use Case

### 1. Pricing Strategy Optimization

Real estate professionals and sellers can leverage the regression model findings to refine their pricing strategies. Understanding the influential predictors, such as Building Square Feet and Land Acre, allows for more accurate property valuation, potentially leading to optimized listing prices.

### 2. Buyer Decision Support

Prospective homebuyers can benefit from the insights into influential predictors when considering property purchases. The information can guide them in evaluating the relative importance of features and making informed decisions aligned with their preferences and budget constraints.

### 3. Investment Decision-Making

Investors in the North Chicago Township real estate market can use the analysis to identify key predictors driving property values. This information aids in making strategic investment decisions, such as identifying properties with potential for appreciation or areas with promising growth.

### 4. Marketing Strategies

Real estate agents can tailor their marketing efforts based on the factors identified as influential in determining property prices. Highlighting features such as larger living spaces or extensive property lots in promotional materials can enhance property desirability.

### 5. Risk Mitigation

Understanding the impact of predictors like Age on property prices can assist stakeholders in assessing potential risks associated with older properties. This insight is valuable for property appraisers, inspectors, and insurers in evaluating risk and determining appropriate pricing or coverage.

### 6. Policy and Planning

Local authorities and policymakers can utilize the information to inform urban development plans and housing policies. Insights into predictors like Tract Median Income can contribute to more targeted and equitable planning efforts.

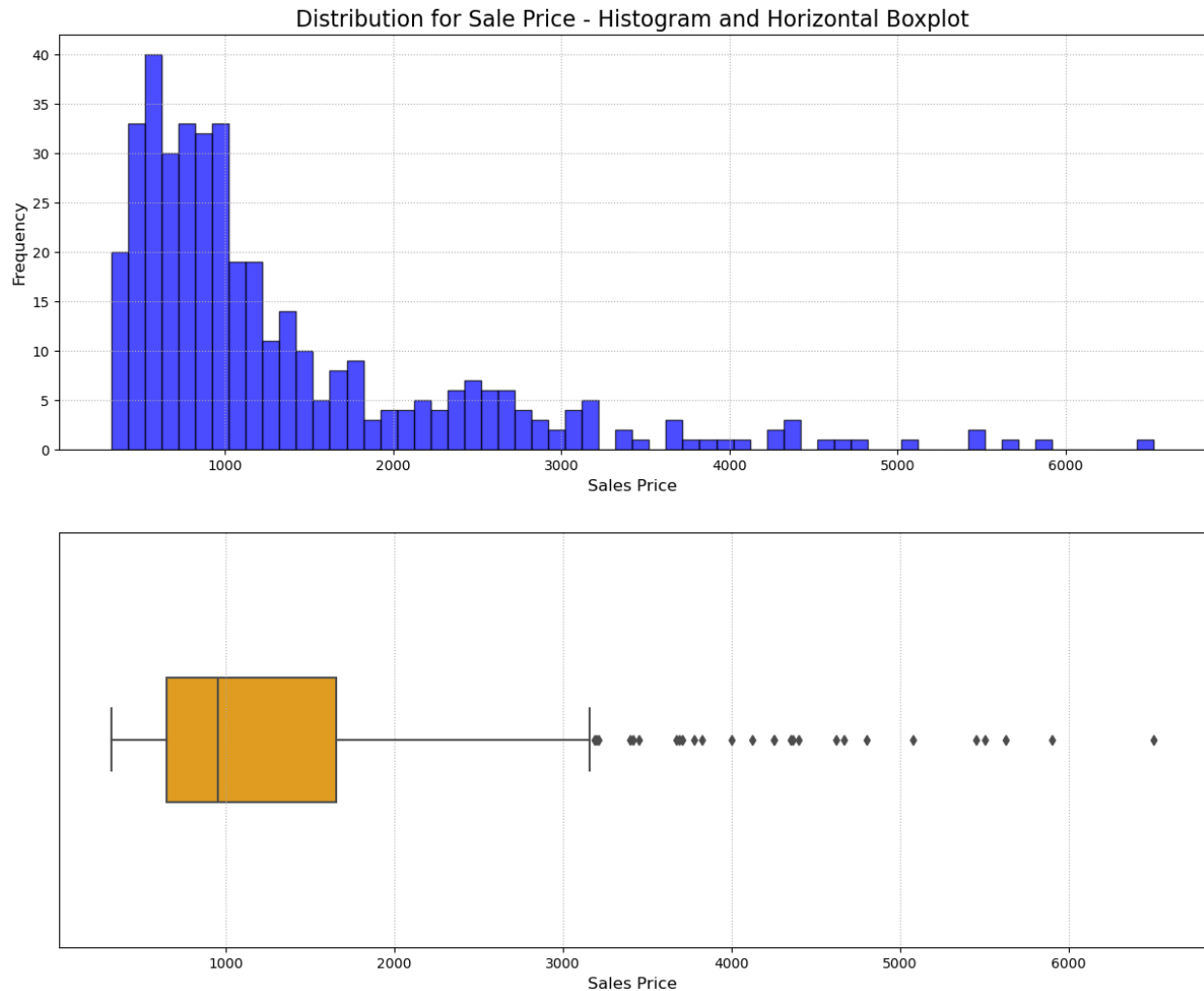
### 7. Educational Resources

Providing homebuyers and sellers with educational resources based on the analysis fosters a more transparent and informed real estate market. It empowers individuals with the knowledge needed to navigate transactions confidently.

## **IV. Explorative Data Analysis**

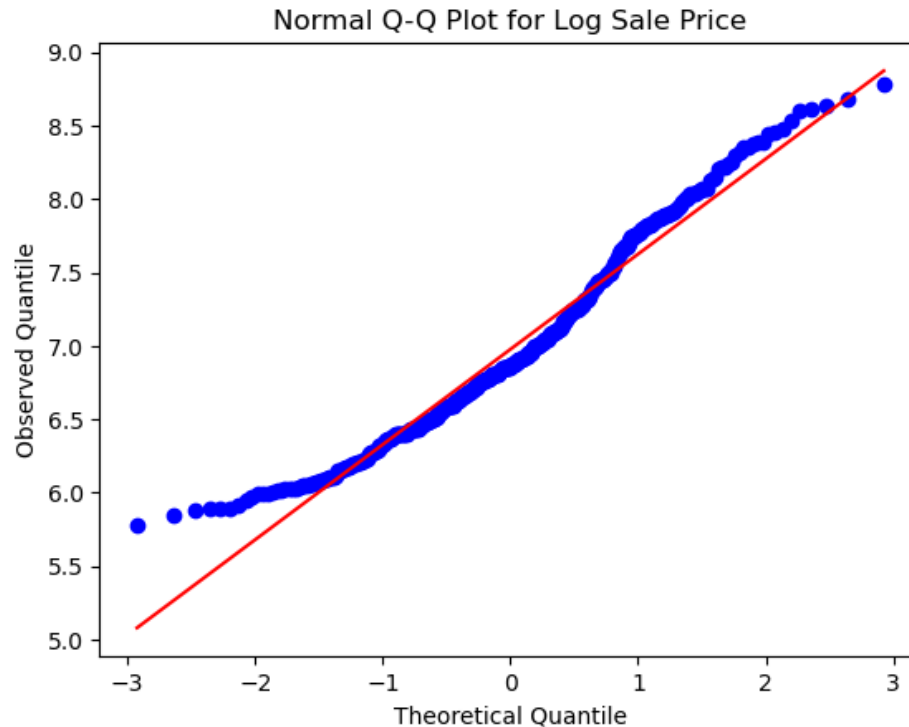
### Data Distribution

We initially examined the distribution of Sale Price. By drawing a histogram with a bin width of 100 and a horizontal boxplot, it revealed a right-skewed distribution, indicative of a non-normal pattern. Instead, it looks more like a Gamma distribution. Since a linear regression model assumes a normal distribution for the error, we will transform the Sale Price to produce a seemingly normal distribution. Many transformations serve this purpose. For us, we will apply the natural logarithm transformation to Sale Price.



With this in mind, we will be checking whether the Log Sale Price is normally distributed. The Normal Q-Q plot, a visual diagnostic tool, did not display the anticipated straight-line pattern indicative of a normal distribution.

Subsequently, formal statistical tests, the Shapiro-Wilks test, and the Anderson-Darling test, were employed for quantitative evaluation. The Shapiro-Wilks test yielded a p-value of 0.9655885, which is considerably high, indicating a lack of evidence to reject the null hypothesis that the data follows a normal distribution. However, the Anderson-Darling test statistic of 4.041554216991244 exceeded the critical values of [0.57, 0.65, 0.779, 0.909, 1.081], leading to the rejection of the null hypothesis.



- Shapiro-Wilks test p-value: 0.9655885
- Anderson-Darling test statistic: 4.041554216991244
- Anderson-Darling test critical values: [0.57 0.65 0.779 0.909 1.081]

However, even if the Log Sale Price does not perfectly follow a normal distribution, it can still be acceptable for linear regression. Linear regression is robust, and some deviations from normality are often tolerable, especially with larger sample sizes. It's important to consider the overall context, the assumptions of linear regression, and whether the model's performance is satisfactory.

### Multiple Linear Regression Model

Next, we transitioned to building a multiple linear regression model incorporating eight predictors, including Age, Bedrooms, Building Square Feet, Full Baths, Garage Size, Half Baths, Land Acre, and Tract Median Income. The model, inclusive of an Intercept term, revealed a commendable Coefficient of Determination (R-squared) of 0.8384, signifying a strong explanatory power.

Diving deeper into the model's intricacies, we examined the regression coefficients along with their standard errors and 95% confidence intervals. The coefficients illuminated the impact of each predictor on Log Sale Price. Notably, variables such as Land Acre and Building Square Feet demonstrated substantial influence, signifying that a larger living space and property lot contribute positively to the property's perceived value. On the other hand, Age exhibited a negative impact, suggesting that older properties, on average, tend to command lower prices in the North Chicago Township real estate market.

These nuanced findings highlight the complex interplay of various factors influencing property prices. Stakeholders, such as homebuyers, sellers, and real estate professionals, can leverage this information to make more informed decisions and refine their strategies in the dynamic North Chicago Township real estate landscape.

#### OLS Regression Results

Dep. Variable:	Log Sale Price	R-squared:	0.838
Model:	OLS	Adj. R-squared:	0.835
Method:	Least Squares	F-statistic:	255.4
Date:	Mon, 04 Dec 2023	Prob (F-statistic):	9.39e-151
Time:	17:20:46	Log-Likelihood:	-33.864
No. Observations:	403	AIC:	85.73
Df Residuals:	394	BIC:	121.7
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	5.5911	0.097	57.795	0.000	5.401	5.781
Age	-0.0045	0.001	-3.005	0.003	-0.007	-0.002
Bedrooms	0.0547	0.020	2.684	0.008	0.015	0.095
Building Square Feet	0.1699	0.023	7.422	0.000	0.125	0.215
Full Baths	0.1122	0.022	5.194	0.000	0.070	0.155
Garage Size	0.0222	0.027	0.812	0.417	-0.032	0.076
Half Baths	0.0561	0.027	2.083	0.038	0.003	0.109
Land Acre	6.8474	0.870	7.874	0.000	5.138	8.557
Tract Median Income	0.0017	0.000	3.648	0.000	0.001	0.003

Omnibus:	22.581	Durbin-Watson:	2.031
Prob(Omnibus):	0.000	Jarque-Bera (JB):	34.493
Skew:	-0.409	Prob(JB):	3.24e-08
Kurtosis:	4.176	Cond. No.	8.32e+03

#### Sale Price Prediction

Finally, we will predict the Sale Price of a single-family home whose features are at the median of all predictors. The calculated predicted Sale Price was found to be \$886.03. This estimate was derived using a multiple linear regression model, taking into account the medians of all predictor variables. Furthermore, the 95% confidence interval for the predicted Log Sale Price was computed, and upon exponentiation, the confidence interval for Sale Price was obtained. The resulting interval, ranging from \$844.95 to \$929.10, provides a measure of uncertainty around the predicted Sale Price.

To conclude this project, we will calculate the Shapley values for the predictors included in our regression model. The Shapley values, representing the average contribution of each predictor to all possible model combinations, were calculated. The sum of Shapley values was found to be 0.8384, indicating the collective influence of all predictors in explaining the variability in sale prices.

Analyzing the individual Shapley values for each predictor, it becomes evident that "Building Square Feet" emerges as the most influential feature, making the highest contribution to predicting sale prices. Following closely are "Land Acre" and "Full Baths," contributing

significantly to the overall model. Other predictors, such as "Bedrooms," "Garage Size," "Age," "Half Baths," and "Tract Median Income," also play roles in determining sale prices, albeit to varying degrees.

Feature	Shapley Values	Percent Shapley
Building Square Feet	0.1955317	23.32%
Land Acre	0.1593620	19.01%
Full Baths	0.1503659	17.94%
Bedrooms	0.1240584	14.80%
Garage Size	0.0917867	10.95%
Age	0.0653106	7.79%
Half Baths	0.0393081	4.69%
Tract Median Income	0.0126395	1.51%

## V. Insight and Conclusion

The exploration of the North Chicago Township real estate dataset has uncovered pivotal insights guiding our understanding of property prices in this dynamic market:

Initial analyses revealed a right-skewed distribution of Sale Prices, prompting a transformation to Log Sale Price for alignment with regression model assumptions.

The ensuing multiple linear regression model, incorporating eight predictors, demonstrated robust performance with a high Coefficient of Determination (R-squared) of 0.8384.

Building Square Feet emerged as the most influential predictor, followed by Land Acre and Full Baths, as confirmed by both regression coefficients and Shapley values.

Utilizing median values, we predicted a Sale Price of \$886.03, supported by a confidence interval providing a nuanced range for anticipated property values.

These findings collectively empower stakeholders with actionable insights, whether optimizing pricing strategies, understanding predictor impacts, or navigating the intricacies of the North Chicago Township real estate market.