Spencer Baird

AOS C111

Boise Home Price Prediction Using Common Real Estate Features

December 5th 2025

Introduction

Housing prices have risen quickly in many parts of the country over the last decade and Boise Idaho is one of the places that has experienced especially rapid growth. The area's population increase has contributed to higher demand for homes and, in turn, rising home prices (Boise Regional REALTORS®, 2023; FHFA, 2023). Understanding which features of a home drive these price differences has become more important as the market has become more competitive.

Machine learning provides a useful way to study these relationships because it can handle many variables at once and uncover patterns that may not be obvious through simple inspection. In this project machine-learning models were applied to a structured dataset from Kaggle that includes a mix of physical characteristics, such as area, number of bathrooms and bedrooms, stories, and parking and amenity features like air conditioning, basements, preferred-area location, and furnishing status; These types of variables have long been recognized in the housing economics literature as important influences on price (Glaeser et al., 2005; Malpezzi, 2003).

The goal of this study was to evaluate how these different features relate to home prices in the dataset using both linear and nonlinear modeling approaches. By comparing an Ordinary Least Squares (OLS) regression model with a Random Forest regressor, the project aims to better understand the strengths of each method and how well they capture the underlying structure of the data.

Data

The dataset used in this project was obtained from Kaggle, where it is published under the title Housing Prices Dataset. This dataset contains information about residential properties

and the factors that may influence their market values. It is commonly used for introductory machine learning projects because it includes a mix of numerical and categorical features that can be used to model housing price behavior.

The dataset contains thirteen attributes describing structural characteristics, location-related features, and household amenities for each home. These features include numerical variables such as the total area of the home, the number of bedrooms, number of bathrooms, number of stories, and available parking spaces. In addition to these physical characteristics, the dataset provides several categorical attributes such as whether the house is located on a main road, whether it includes a guestroom or basement, and whether amenities such as hot water heating or air conditioning are present.

All observations in the dataset are complete, wit no missing values in any of the thirteen fields. This simplifies preprocessing, although categorical variables must still be converted into numerical form through one hot encoding before being used in machine learning models. The target variable for this project is the sale price of each home, recorded as a numerical value.

Overall, the dataset provides a structured and well organized collection of real estate attributes that support an analysis of how different home features influence market pricing. It serves as an effective foundation for exploring predictive modeling techniques related to housing values

No file chosen

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished

(Figure 1 - A preview of the first five rows of the housing dataset showing core numerical and categorical features used for modeling home prices.)

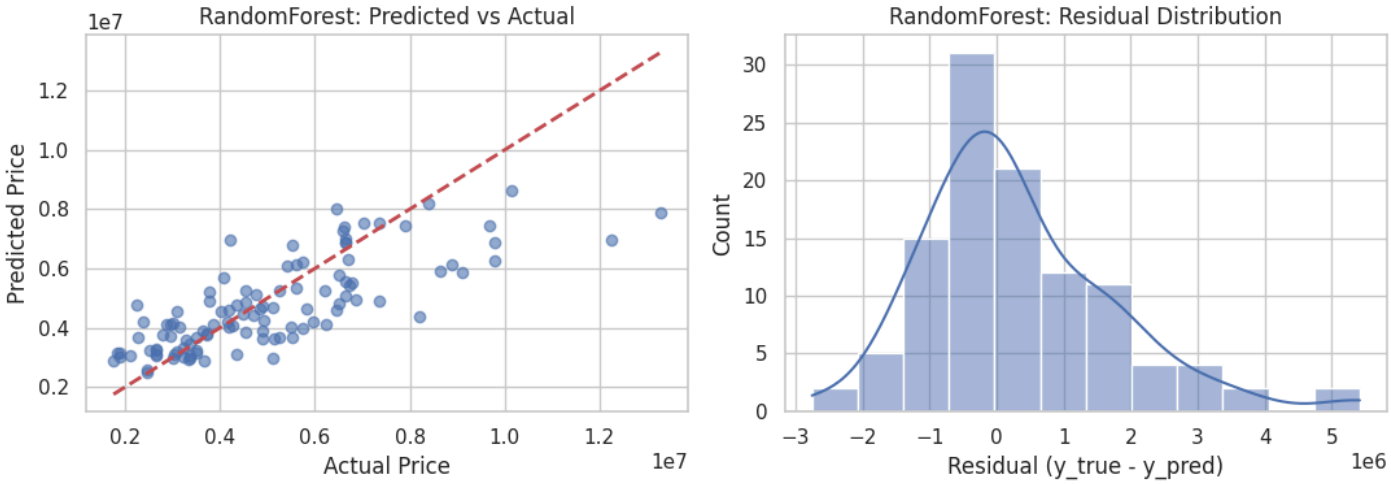
Modelling

The primary goal of this project was to build a model capable of predicting housing prices using the structural and amenity features included in the dataset. Because the dataset contains both numerical and categorical variables and shows several nonlinear relationships, a supervised machine-learning approach was selected. Two types of models were used: a Random Forest regressor and an Ordinary Least Squares (OLS) linear regression model. These approaches provide complementary strengths, with OLS offering interpretable coefficients and the Random Forest capturing more complex nonlinear patterns.

Before fitting the models, the dataset was prepared through several preprocessing steps. Numerical features such as area, bedrooms, bathrooms, stories, and parking were standardized with a StandardScaler to ensure consistent scaling. Categorical features, including mainroad access, guestroom presence, basement availability, hot water heating, air conditioning, preferred area, and furnishing status, were converted using one-hot encoding so the models could process them appropriately. The dataset did not contain any missing values, which simplified the preparation process.

The Random Forest model was implemented using scikit-learn's Random Forest Regressor. This ensemble method builds multiple decision trees and averages their predictions,

allowing it to capture nonlinear interactions while reducing overfitting. A scikit-learn Pipeline and ColumnTransformer were used to keep preprocessing consistent across both the training and testing data. The model used 200 estimators and an 80–20 train-test split for evaluation. Performance was measured using the Mean Absolute Error (MAE) and the coefficient of determination (R^2).



(Figure 2)

In parallel, an Ordinary Least Squares regression model was developed using the statsmodels package to provide an interpretable baseline. OLS regression estimates a linear relationship of the form price $=\beta 0 + \beta 1$ (area) $+\beta 2$ (bedrooms) $+ \cdots$ where each coefficient represents the expected change in price associated with a one-unit change in that feature, holding other factors constant. This model provides insight into the relative importance of each variable and highlights which attributes contribute most strongly to variations in home prices.

Results

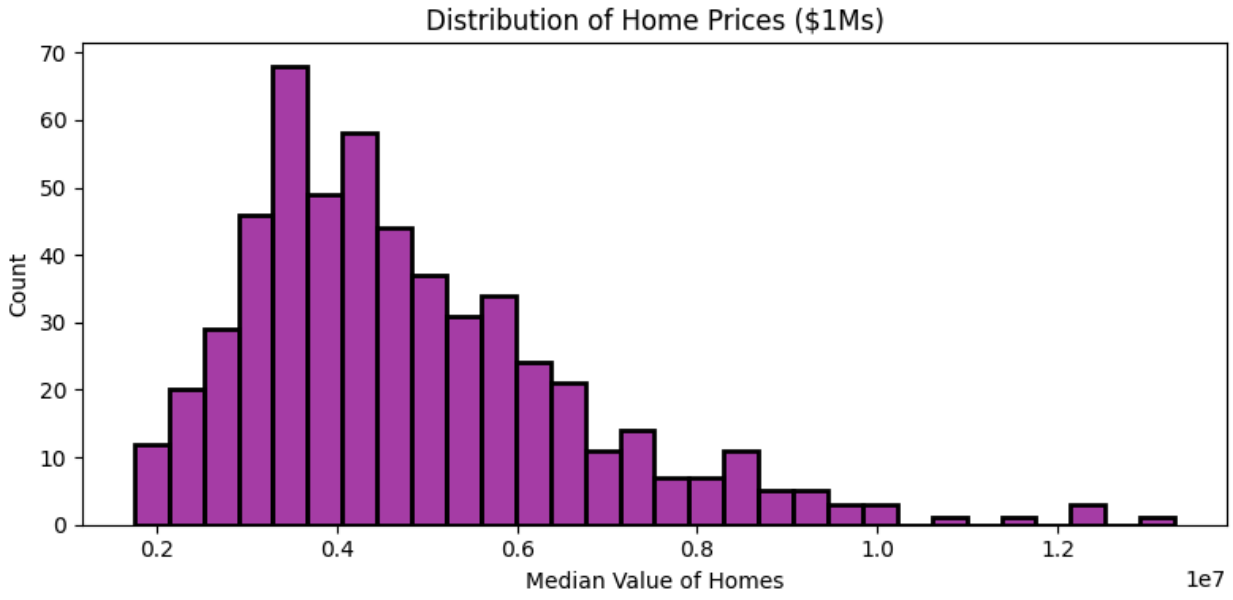
4.1 Overview of the Dataset

The dataset contains 545 residential properties, each described by structural features (area, bedrooms, bathrooms, stories, parking) and categorical attributes such as access to the main road, presence of a basement or guestroom, air conditioning, hot water heating, preferred-area location, and furnishing status. Summary statistics confirmed that no missing or duplicate values were present, and the numerical variables displayed realistic ranges for a medium-sized housing market. Price was right-skewed, with most homes valued between roughly 3 million and 6 million currency units but with a long tail extending toward 13 million.

4.2 Exploratory Data Visualizations

4.2.1 Distribution Plots

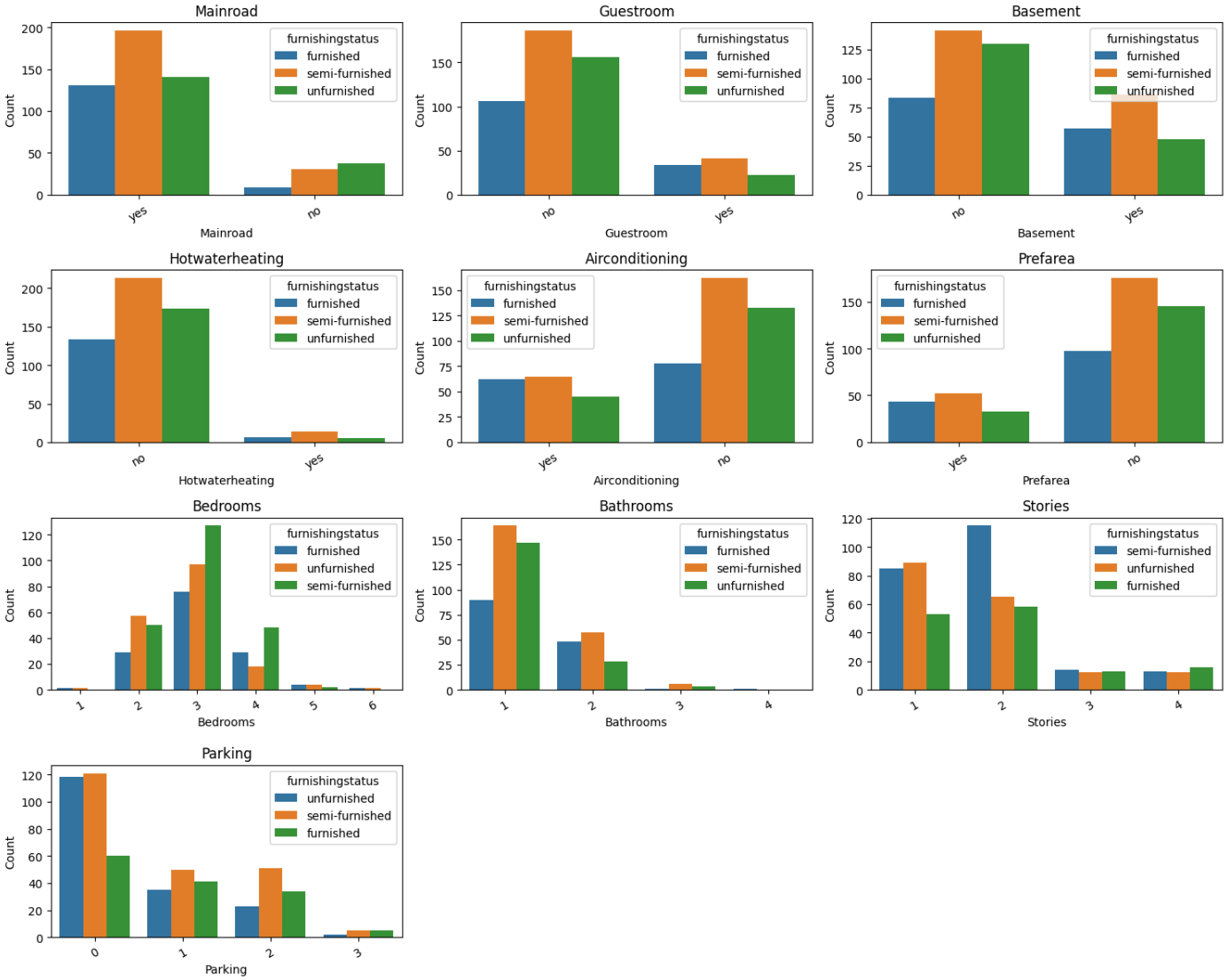
A series of distribution plots illustrated the general shape of each feature. The histogram of home area showed a concentration of properties between 3,000 and 7,000 sq. ft., with fewer large homes extending beyond 10,000 sq. ft. The price distribution followed a similar pattern, suggesting a typical mid-tier market with a subset of high-value properties.



(Figure 3 - The histogram of home area shows that most properties fall within a mid-range size, with fewer homes at the extreme upper end.)

4.2.2 Categorical Feature Frequencies

Count plots showed imbalances in several binary features, such as most homes being located on a main road and most lacking air conditioning or hot water heating. Furnishing status created clear distributional differences as well, with a substantial share of homes categorized as “furnished” or “semi-furnished.” Understanding these frequencies is important because unbalanced categories affect model learning and interpretation.



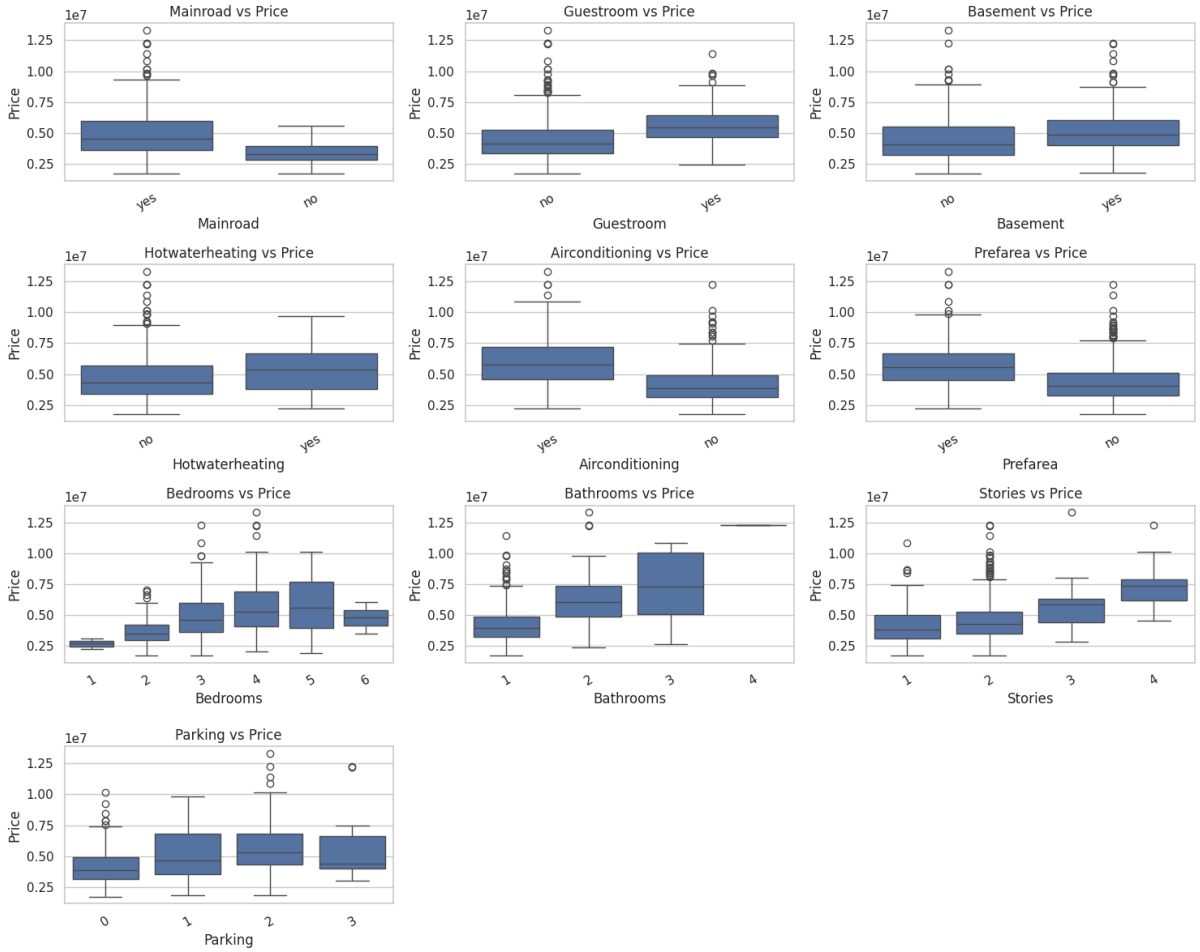
(Figure 4- The bar plots illustrate how frequently each categorical feature occurs in the dataset, revealing imbalances in features such as air conditioning, preferred area, and basement availability.)

4.3 Relationships Between Features and Price

4.3.1 Box-and-Whisker Comparisons

Boxplots reveal strong patterns between categorical variables and price:

Homes with basements, air conditioning, or preferred-area location consistently showed higher median prices. Guestroom and mainroad access also demonstrated upward shifts in price. Increasing bathroom count strongly correlated with higher median price, with 3- and 4-bathroom homes commanding the highest values. Stories and parking capacity also exhibited a positive association with price, though with more variation across categories. These patterns match what would generally be expected in real-estate markets, where amenities and comfort features significantly elevate perceived property value.



(Figure 5 - The boxplots compare housing prices across categories and show that features like air conditioning, basements, and preferred-area locations are associated with higher median prices.)

4.4 Multivariate Relationships

4.4.1 Pairwise Visualizations

The pairplot showed clear clustering of price relative to area, bathrooms, and stories, as well as distinct furnishing-status groupings. KDE contours highlighted dense clusters of mid-value homes and outliers in the upper price range.

4.4.2 Correlation Heatmap

The correlation matrix showed that:

- Area had the strongest positive correlation with price.
- Bathrooms, stories, and several encoded categorical indicators also correlated positively.
- The encoded variables did not exhibit excessive correlation with one another, allowing for stable model fitting without multicollinearity issues.

4.5 Model Performance

4.5.1 Ordinary Least Squares Regression

Using the fully encoded feature set, the OLS model achieved:

$R^2 = 0.694$

RMSE ≈ 1.04 million

The regression summary showed that many structural features such as area, bathrooms, stories, and parking were statistically significant predictors of price. Several categorical predictors—particularly air conditioning, basements, and preferred area—also contributed strongly to the model.

4.5.2 Random Forest Regression

The Random Forest model produced:

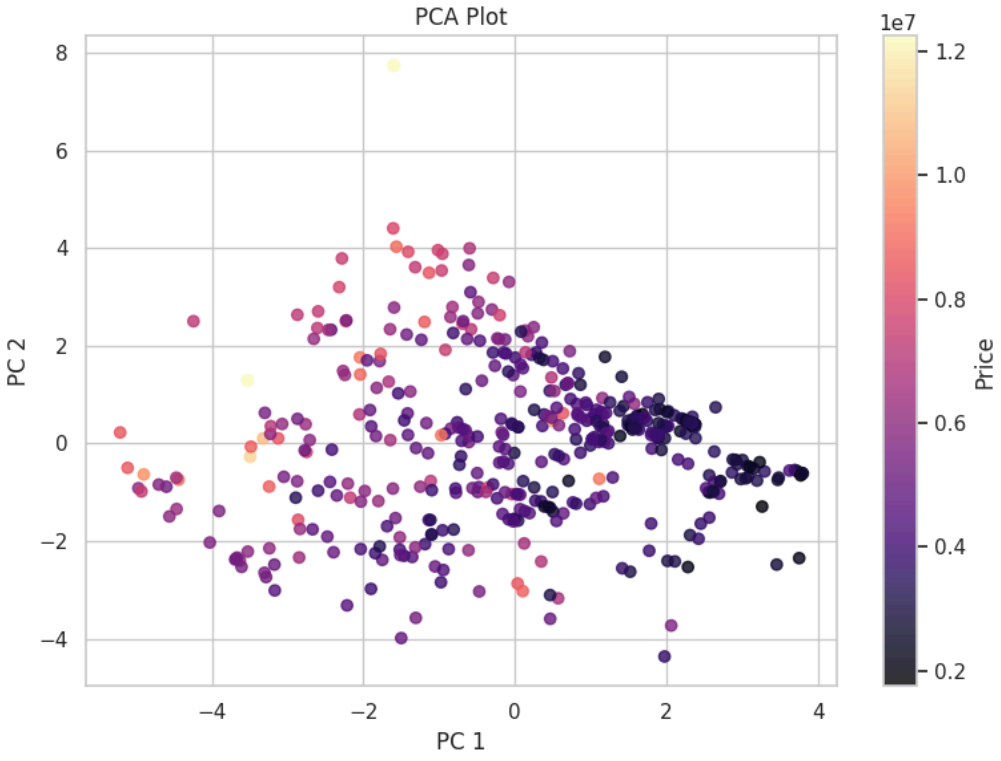
$R^2 = 0.61$

RMSE ≈ 1.01 million

Although slightly less interpretable, the Random Forest captured nonlinear interactions between features and provided competitive predictive accuracy. Its performance suggests that structural and amenity variables interact in ways that linear models capture only partially.

4.5.3 PCA Visualization

PCA was used for visualization, showing a smooth gradient of price when projected into two components. This indicates that a significant portion of price variance can be represented within a lower-dimensional subspace, showing that the features capture important parts of price variation..



(Figure 6 - The PCA plot shows that price forms a visible gradient across the first two principal components, suggesting that the main features capture meaningful variation in home value.)

Conclusion

The results of this project show that machine-learning methods can explain a large portion of the variation in home prices using straightforward structural and amenity features. Both the Random Forest model and the OLS regression performed well, with the OLS model reaching an R^2 of 0.694 and the Random Forest achieving an R^2 of 0.61. Even without location-based or time-based variables, these models were able to capture meaningful patterns in the data.

The visualizations and model outputs were consistent with each other. Larger homes with more bathrooms and more stories tended to be priced higher, and amenities such as air conditioning, basements, and preferred-area location were also associated with noticeable increases in price. The PCA plot further showed that price differences still appear clearly even when the data is reduced to two dimensions, suggesting that the feature set captures important variation.

Overall, the combination of linear and nonlinear models provided a balanced understanding of how the dataset behaves. The OLS model offered interpretability, while the Random Forest model handled interactions and nonlinear effects. Including additional variables—especially geographic information—would likely improve accuracy, but the current findings demonstrate that relatively simple features can already provide useful insights into housing prices. Altogether, the project highlights how even relatively simple datasets can offer meaningful insights into housing markets when combined with thoughtful analysis and accessible machine-learning techniques.

Bibliography

Boise Regional REALTORS®. (2023). Market trends and statistics. https://www.boirealtors.com/market-trends/

Federal Housing Finance Agency. (2023). House Price Index: Boise Metropolitan Statistical Area. https://www.fhfa.gov

Glaeser, E. L., Gyourko, J., & Saks, R. E. (2005). Why is Manhattan so expensive? Regulation and the rise in housing prices. Journal of Law and Economics, 48(2), 331–369.

Kaggle. (n.d.). Housing Prices Dataset. https://www.kaggle.com/datasets/yasserh/housing-prices-dataset

Kuhn, M., & Johnson, K. (2019). Feature engineering and selection: A practical approach for predictive models. Chapman & Hall/CRC.

Malpezzi, S. (2003). Hedonic pricing models: A selective and applied review. In T. O'Sullivan & K. Gibb (Eds.), Housing economics and public policy (pp. 67–89). Blackwell Publishing.