Introduction

The housing market in Boise, Idaho has changed rapidly over the past decade, shifting from a relatively affordable regional city into one of the fastest growing and most expensive markets in the country. Population growth, limited housing supply, migration from higher cost states, and evolving economic conditions have combined to push home prices to historic highs. As someone from Boise, I have watched these changes occur in real time. Many neighborhoods that once offered accessible entry level homes have become challenging for first time buyers, and price instability has become a major concern for residents and planners.

The central problem addressed in this study is the challenge of predicting housing prices using available property features. Although there is a great deal of national level housing research, local markets such as Boise behave in unique ways due to regional demographics, land use patterns, and economic influences. Traditional appraisal methods do not always keep pace with rapid market movement, and they can be limited by subjective judgment. Machine learning approaches allow the analysis of complex relationships among property characteristics and make it possible to generate more consistent and data driven price estimates.

This problem carries particular importance in the Boise region because the area has experienced some of the strongest price growth in the United States. For example, the Federal Housing Finance Agency reported a large increase in median home prices between 2020 and 2021, at one point exceeding forty percent growth in a single year. Although prices cooled somewhat in 2022 and 2023, values remain significantly higher than before the pandemic. Accurate price prediction supports buyers who are trying to make informed financial decisions, real estate professionals who need dependable valuations, and policymakers who are working to address concerns about housing affordability. By applying modern machine learning methods to a structured housing dataset, this project seeks to improve understanding of the main factors influencing home values and to contribute a data informed perspective on the continuing evolution of the Boise housing market.

We did this to solve the problem. We concluded that...

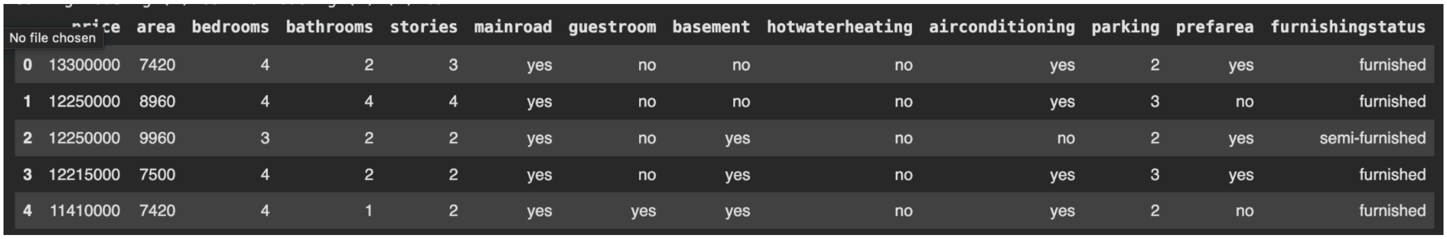
Data

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. Location preference is captured through a variable indicating whether the home is situated in a preferred area, and the furnishing status of each home is categorized as furnished, semi furnished, or unfurnished.

All observations in the dataset are complete, with no missing values in any of the thirteen fields housing_price_project.ipynb. - …. 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 develop a model capable of predicting housing prices based on structural characteristics and amenity features included in the dataset. Because the dataset contains a combination of numerical and categorical variables and exhibits multiple nonlinear relationships among features, a supervised machine learning approach was chosen. Two modeling strategies were explored: a Random Forest regression model and an Ordinary Least Squares (OLS) linear regression model. These complementary approaches allow both interpretable coefficient-based analysis and more flexible nonlinear prediction.

Before applying these models, several preprocessing steps were performed. The numerical features, including area, bedrooms, bathrooms, stories, and parking, were standardized using a StandardScaler to ensure consistent scaling across variables. The categorical features, such as mainroad access, guestroom presence, basement availability, hot water heating, air conditioning, preferred area, and furnishing status, were transformed using one-hot encoding. This allowed the models to handle non-numeric information by converting each category into a binary indicator variable. The dataset contained no missing values, simplifying preprocessing.

The Random Forest model was implemented using the RandomForestRegressor class from the scikit-learn library. This ensemble method constructs multiple decision trees and averages their predictions, enabling the model to capture nonlinear interactions between features and reduce overfitting. The training and evaluation pipeline was constructed using the scikit-learn Pipeline and ColumnTransformer classes, which ensured that all preprocessing steps were applied consistently to both the training and testing data. The model was trained using two hundred estimators and evaluated using an 80–20 train-test split. Model performance was assessed using the mean absolute error (MAE) and the coefficient of determination (R^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

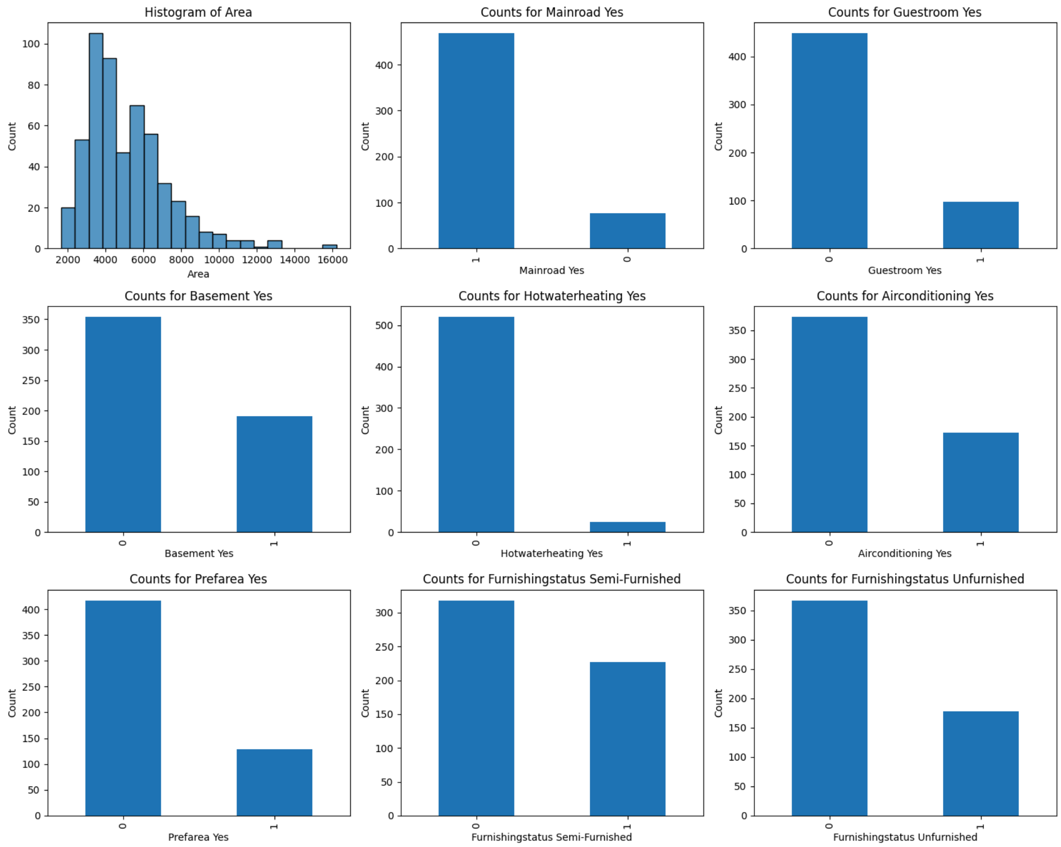
$\text{price} = \beta 0 + \beta 1 \text{ (area) } + \beta 2 \text{ (bedrooms) } + \cdots$

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.

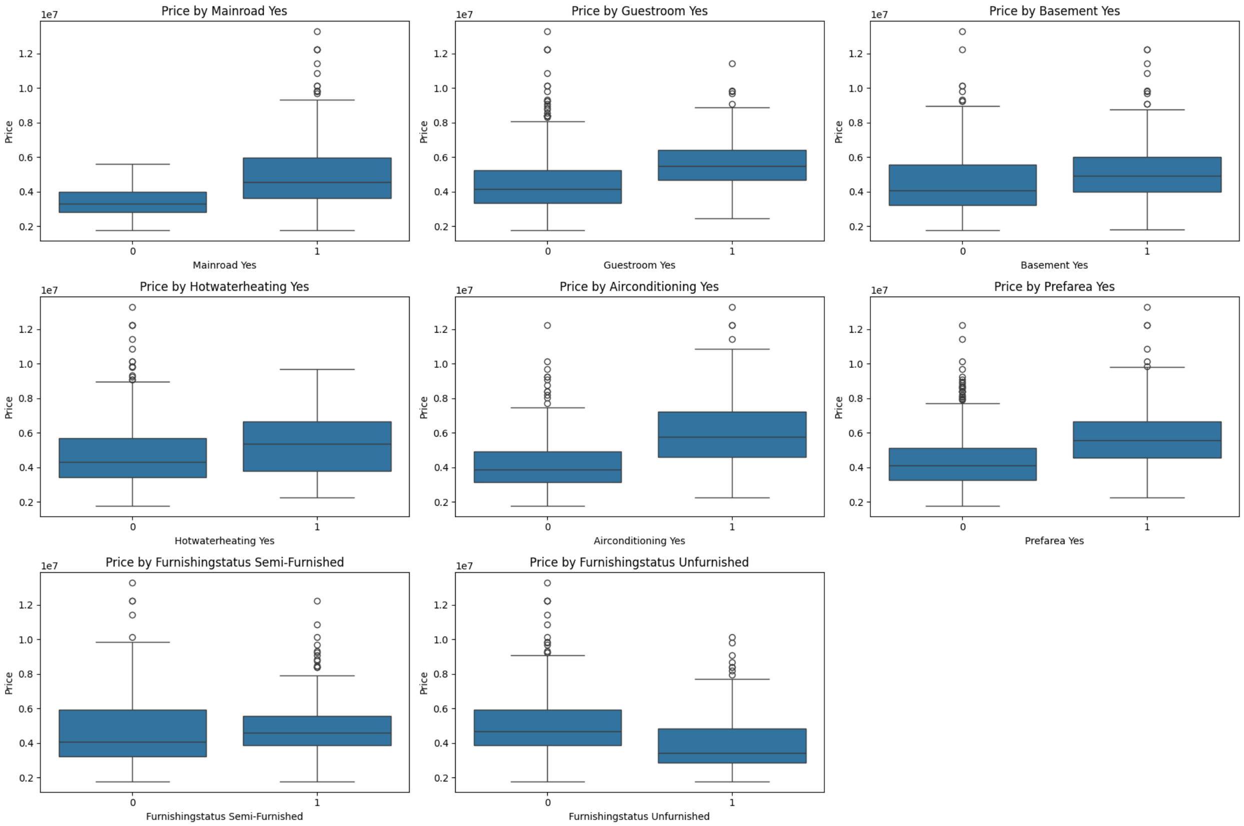
An example of the model-fitting code is shown below:

Results

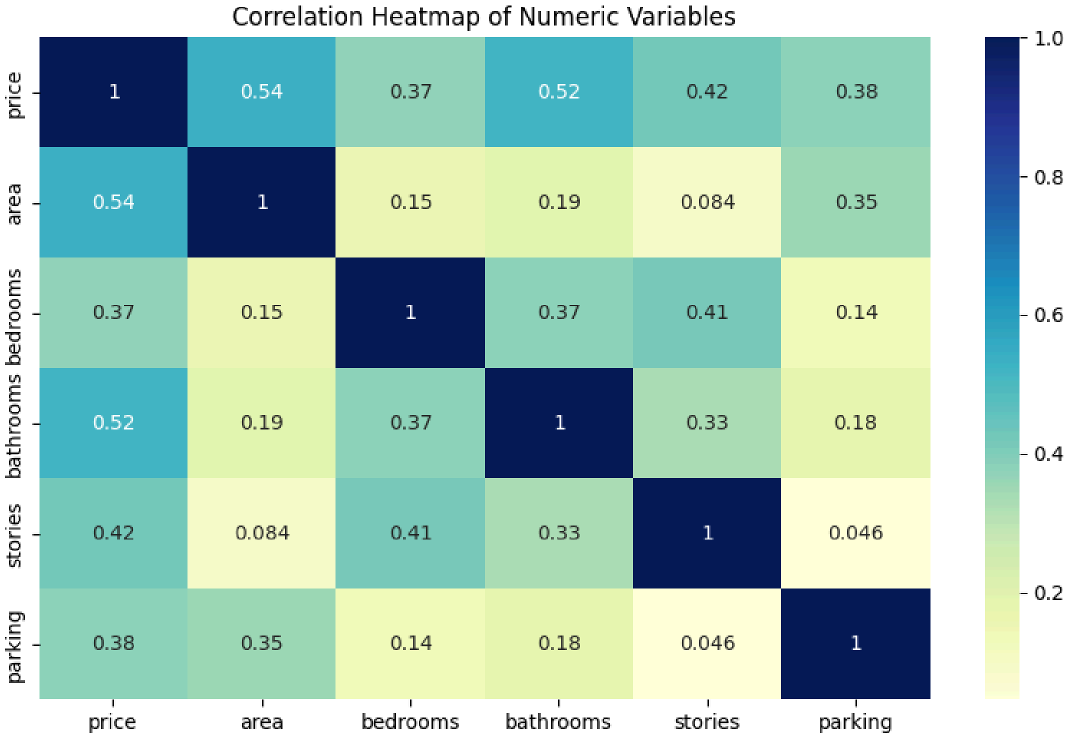
The Random Forest model performed reasonably well on the test set, achieving a mean absolute error (MAE) of approximately 1.01 million and an R² score of 0.61, indicating that the model was able to explain about 61 percent of the variance in home prices. The OLS regression model performed slightly better from an explanatory standpoint, with an R² of 0.68, suggesting that a substantial portion of price variability can be captured through a linear combination of structural and amenity variables. Taken together, these results show that both models are capable of identifying meaningful patterns in the housing data, with the Random Forest providing flexible nonlinear predictions and the OLS model offering strong interpretability.



(Figure 2 Canvas of Numerical Historgram and Catagorical Box Plot)



(Figure 3 Canvas of Box and Whisker plots comparing price to categorical data)



	price	area	bedrooms	bathrooms	stories	parking
price	1	0.54	0.37	0.52	0.42	0.38
area	0.54	1	0.15	0.19	0.084	0.35
bedrooms	0.37	0.15	1	0.37	0.41	0.14
bathrooms	0.52	0.19	0.37	1	0.33	0.18
stories	0.42	0.084	0.41	0.33	1	0.046
parking	0.38	0.35	0.14	0.18	0.046	1

(Figure 4 Correlation Heat Map of Numeric Variables)

The exploratory analysis revealed several meaningful patterns in the housing dataset. The combined canvas of histograms and categorical bar charts (Figure 2) provides an overview of the distribution of property characteristics. The histogram of area shows that most homes fall between approximately 3,000 and 7,000 square feet, with a smaller number of very large homes represented in the upper tail. The categorical features, such as mainroad access, basement availability, and furnishing status, show uneven distributions, with some features occurring far more frequently than others. These imbalances help explain some of the variation observed in home prices.

To examine how categorical variables relate to home prices, a series of box-and-whisker plots was generated (Figure 3). Several patterns emerge clearly from these visualizations. Homes with air conditioning, basements, and preferred-area locations consistently show higher median prices than homes without these features. Similarly, properties located on the main road or containing a guestroom exhibit modestly elevated price distributions. The furnishing status variable shows notable separation as well, with fully furnished homes tending to command higher prices than semi-furnished or unfurnished properties. These patterns align with expectations for a dataset where comfort amenities and location advantages contribute meaningfully to market value.

A correlation heatmap of the numerical variables (Figure 4) provides additional insight into linear relationships within the dataset. Price shows the strongest correlation with area and number of bathrooms, indicating that larger homes and those with more amenities tend to sell for higher amounts. Stories and parking also exhibit moderate positive correlations with price, while the number of bedrooms shows a weaker relationship. The correlations among the predictors themselves are generally moderate, suggesting that the features capture distinct aspects of home characteristics.

Conclusion

This project demonstrated the usefulness of machine learning techniques for predicting housing prices based on structural characteristics and amenity-related features. Both the Random Forest and the OLS regression models identified clear and interpretable relationships within the dataset. The Random Forest achieved an R^2 score of 0.61 with a mean absolute error of approximately 1.01 million, while the OLS model performed slightly better from an explanatory perspective, reaching an R^2 of 0.68. These results show that a significant portion of the variation in home prices can be captured using the available features, even within a relatively compact dataset.

The exploratory visualizations revealed consistent trends: homes with larger areas, more bathrooms, or comfort amenities such as air conditioning and basements tended to command higher prices. Preferred-area location and furnishing status also played notable roles in shaping price distributions. These observations align with typical patterns seen in real housing markets, including high-growth regions such as Boise, where home size, amenities, and neighborhood desirability are major price drivers.

Overall, the findings indicate that the dataset provides a solid foundation for predictive modeling and that both linear and nonlinear approaches can extract meaningful insights into the determinants of housing value. While more complex models or additional geographic and economic variables could further enhance predictive performance, the models developed here offer a clear and practical demonstration of how machine learning can support housing market analysis and price estimation.