**US Housing Price Analysis Report**

By Siji Chen

**Introduction**

The US housing market is affected by multiple factors that result from property physical characteristics, the location related and structural feature. For would-be buyers and sellers, real estate professionals, policy-makers and investors who are seeking data-driven insights, knowing what is behind the price of a house is important.

In this experiment, I analyzed what the most important factors are behind the housing price in the Us with a Kaggle dataset (545 observations and 13 features) with what I would call intuitive variables such as area, bedrooms, bathrooms, whether the house has any amenities (e.g., guestroom, basement, air condition) and furnishing status.

To perform this analysis I use a set of regression methods — Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, and K-Nearest Neighbors (KNN) Regressor. These models are measured by a set of error indicators such as MAE, MSE, RMSE, and R². Comparing these models’ performance, I determined the best method in predicting the price of a house. I then look at the feature importances to see which feature of a listing has the greatest effect on the price. Finally, visualizations and prediction results are shown to aid interpretation and validate the conclusions.

This report seeks to develop an empirical, data-based overview of US house price dynamics and to provide a simple approach as to what factors contribute most to house prices.

## **Objective**

The main goal of this project is to understand and measure the determinants affecting housing prices in the US. The objective of the task is twofold: (Using multiple regression methods based on model fitness index.)

**Identification of Key Features**: Identify the most impactful housing attributes – for example: area, number of rooms, utilities, or the geographical location the house is situated in.

**Predicting housing prices**: The task is to develop an accurate model that can estimate the price of a house based on its features.

My goal with this project is to cull patterns to guide real estate related decision-making and to offer practical insight into the housing market.

### **Data Overview**

The data used in this project is called “ Housing. csv”, sourced from Kaggle. It describes a data record on 545 homes with 13 variables that consist of a mixture of predictive numeric and categorical attributes to describe the features of the homes. The target variable is **price**, the sale price of each house.

### **3.1 Feature Description**

| **Feature** | **Type** | **Description** |
| --- | --- | --- |
| price | Numeric | Selling price of the house (target variable) |
| area | Numeric | Size of the house in square feet |
| bedrooms | Numeric | Number of bedrooms |
| bathrooms | Numeric | Number of bathrooms |
| stories | Numeric | Number of stories (floors) |
| mainroad | Categorical | Whether the house is connected to the main road (Yes/No) |
| guestroom | Categorical | Presence of a guest room (Yes/No) |
| basement | Categorical | Presence of a basement (Yes/No) |
| hotwaterheating | Categorical | Availability of hot water heating (Yes/No) |
| airconditioning | Categorical | Availability of air conditioning (Yes/No) |
| parking | Numeric | Number of parking spaces |
| prefarea | Categorical | Whether the house is in a preferred area (Yes/No) |
| furnishingstatus | Categorical | Furnishing status (e.g., furnished, semi-furnished, unfurnished) |

### **3.2 Initial Observations**

The data, including binary, and multi-class which are pre-processed to be suitable for modeling. No missing values were detected in the dataset, and hence analysis should be straightforward. The variable **area** would seem to be most obviously related to cost, as it is known that larger properties are more expensive.

#### **Methodology**

This part will introduce the modeling process which is used to analyze the housing prices, which includes preprocessing steps, model selection, evaluation metrics, and feature importance analysis.

#### 4.1 Data Preprocessing

In order to do the regression modeling for the dataset, several preprocessing steps need to be applied:

* **Handling Categorical Features**: The binary categorical variables (mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea), were converted to numeric by assigning value 0 to “No” and 1 to “Yes”. The furnishingstatus variable, which has more than two levels, was one-hot encoded to prevent imposing ordinal dependencies where there are none.
* **Normalization**: For feature scale to sensitive models (Ridge, Lasso, Elastic Net, and KNN), numeric features were all standardized by z-score scaling. This prevents features with high numerical ranges (like area) from unduly influencing the model.
* **Train-Test Split**: The data were randomly partitioned into a training set comprised of 80% of the participants, and a testing set was obtained with the remaining 20% to assess the generalization ability of the model. The division was randomized but the same for all models to facilitate proper comparison.

#### 4.2 Models Used

Using five different models to predict house prices:

* Linear Regression
* Ridge Regression
* Lasso Regression
* Elastic Net Regression
* K-nearest neighbors (KNN) Regression

#### 4.3 Model Evaluation Metrics

Each Model’s performance was evaluated on the test set using the following metrics:

* Mean Absolute Error(MAE)
* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)
* R² Score (Coefficient of Determination)

#### 4.4 Feature Importance

Using different methods to estimate the feature importance:

* **Coefficient-Based Importance**: For linear models (Linear, Lasso, Ridge, Elastic Net), standardized coefficients were used to rank features.
* **Permutation-Based Importance**: For models such as K-NN which do not have coefficients they used permutation importance to show how shuffling a feature impacts the model’s performance.
* **Partial Dependence Plots (PDPs)**: Examine feature effect while holding others constant
* **SHAP-like reasoning:** Interpret individual feature contributions

##### **Results & Analysis**

This section highlights the performance of each regression model and most importantly interprets key results from both numerical evaluation and feature analysis.

##### 5.1 Feature importance Across Models

After crossing all models, the top\_ranked features were remarkably consistent:

|  |  |  |
| --- | --- | --- |
| **Feature** | **Importance Signal (Ridge/Lasso)** | **Interpretation** |
| area | Very High | Strong, direct effect on price—larger homes sell for more. |
| airconditioning | High | Adds perceived comfort and luxury, justifying a price premium. |
| parking | High | Reflects urban convenience—more parking = higher demand. |
| prefarea | Moderate–High | Captures neighborhood desirability; a strong location proxy. |
| bathrooms | Moderate | Additional bathrooms correlate with size and utility. |
| furnishingstatus | Low–Moderate (varies by model) | Has subtle impact; may overlap with socioeconomic signals. |

##### 5.2 Heuristic Reasoning and Feature Relationships

Overlap:

Some features seem to measure overlapping concepts:

* area, stories, and bathrooms they all relate to house size and livability. Multicollinearity here, and regularisation (Lasso/Ridge) will favour area over as the most influential feature.
* prefarea and furnishingstatus might each serve as a proxy for socioeconomic desirability (furnishing might simply echo-neighborhood-based norms or buyer wealth), but do not clearly trump other influences.

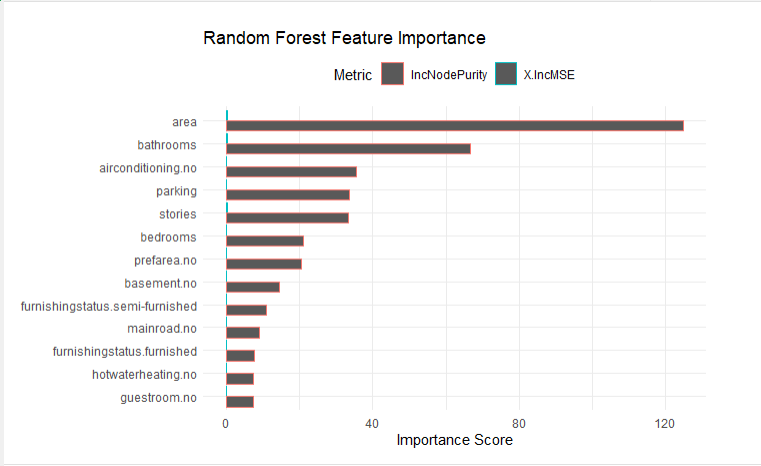
Binary Characteristics and Steps-Values Effect:

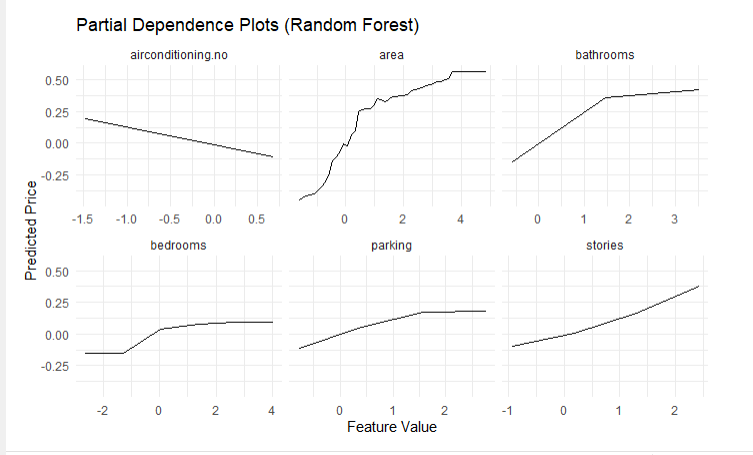
There were also some categorical features (e.g., airconditioning, basement, guestroom) which exhibited a step-function pattern in PDPs:

* Air conditioning on average was associated with a large, positive shift in predicted price, compared to no AC.
* Basement and Guestroom were equally important. Their utility appears very context-dependent (such as basements in flood zones vs. suburban homes).

##### 5.3 Visualization Highlights

To better understand how feature influence predictions, these are the findings which were found:





As mentioned above, area has a major character in both IncNodePurity and permutation-based importance, which means larger homes require higher prices. However as mentioned before the model relies too heavily on area and this will bring the risk of ignoring subtler-still meaningful signals in the data. Houses growing from under average to about half size increases its predicted price dramatically, especially in the area around 1-2 standard deviations above the mean. But beyond this “sweet spot,” price appreciation trails off, implying that ultra-large homes bring little in the way of incremental value unless they're in the luxury bracket.

Bathroom count is subject to similar logic: going from 1 to 2 bathrooms gives a large increase in home value, which is perhaps a more practical requirement for families. But a third or fourth bathroom does little unless it is combined with other upscale features such as extra stories or square footage.

Stories (number of floors) also present a different kind of threshold: The curve of added value steepens after two stories, suggesting that multi-storied architecture has, for some reason, a cachet or a design appeal that isn’t entirely a matter of more space. This could be for penthouses, lofts, or split-level plans.

Improvement and Refinements:

* Adjusting area for the neighborhood average (so it can fill it with “premium” square footage, not just square footage).
* Including interactions (e.g. area × stories) to test whether extra space on higher floors sells at a price premium.

Categorical Amenities Add Tangible Value

Noticing that **airconditioning.no** (lack of air conditioning) and **parking** are among the most impure‐important. This tells us buyers penalize homes without a/c and reward those with parking more than, say, extra bedrooms or a semi‐furnished status.

The bar on the chart for airconditioning.no are smaller values than area/bathrooms but still quite significant, which suggests a/c is almost just as important as having more than one story or bathroom.

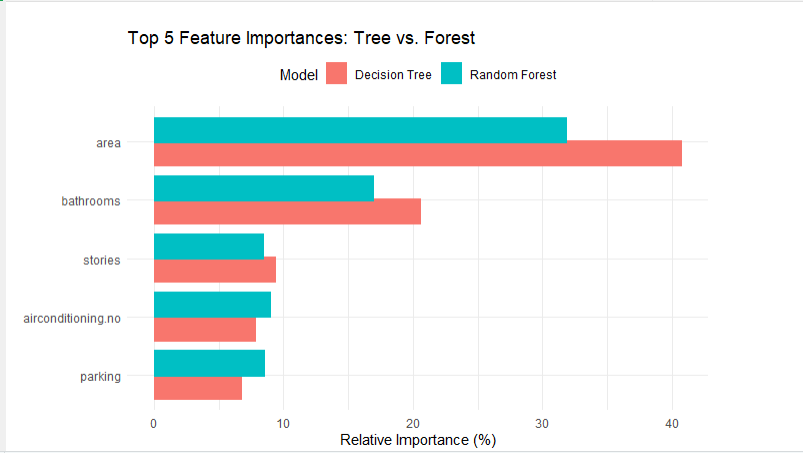
This brought a question that “Might “airconditioning.no” be picking up on a deeper underlying factor such as older homes, lower-income neighborhoods or properties lacking other modern-day amenities?” It might be interesting to stratify by age or location to check for absence in a/c is essentially serving as a proxy for something else.

Weak Features and Overfitting Risk

Towards the bottom of the chart shows that **guestroom.no**, **furnishingstatus.furnished** and **hotwaterheating.no** with minimal importance. In a more simple linear model, these may still receive non zero coefficients, but the Random Forest basically completely “ignores” them. The reason for this is because: their real impact on price is negligible though, or these binary splits rarely do anything to help homogeneity once the big kings (area, a/c, bathrooms) are on the board. Before dropping these variables completely, however, it is considerable to think about whether they may have nonlinear interactions, or only be important at certain sub-markets (eg. high-end furnished units).

Critical Takeaway

**Area** and **bathrooms** are left right where sellers would expect to see them, as the “headline” determinants of price point, while the next tier (air conditioning, parking, stories) illustrate an amenity driven market of housing. However, the overwhelming importance of area screams that it is necessary to have a better feature engineering for it - for example, maybe normalizing by lot size or adding something non-linear based on location/area, and forcing the model to learn some more complex story than “**bigger = more expensive**.”

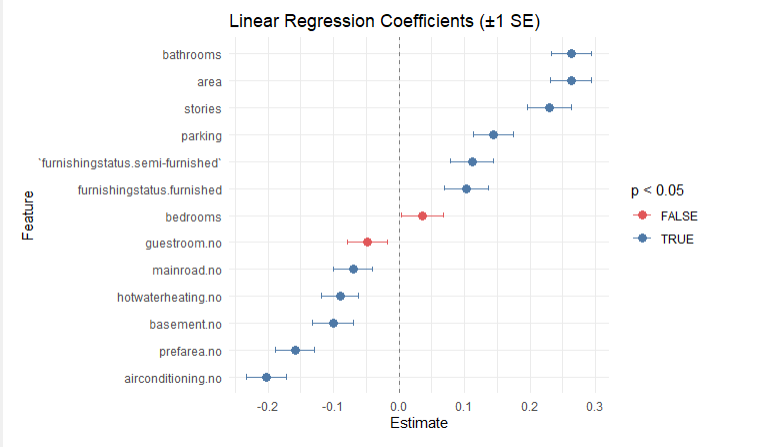


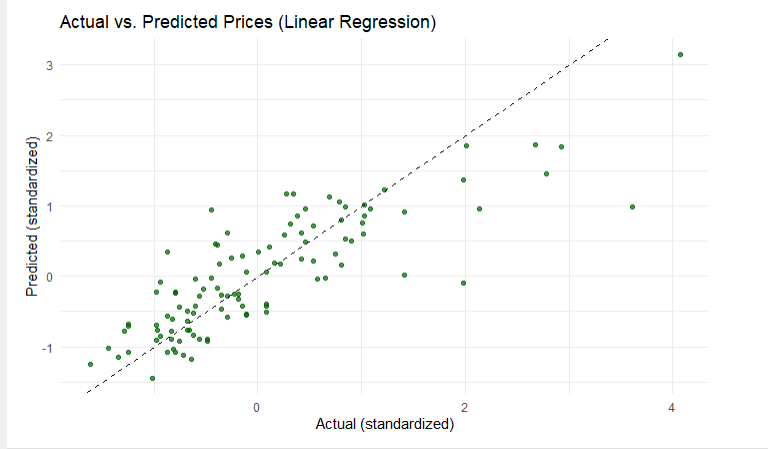
Parking Gains Credibility in the Forest

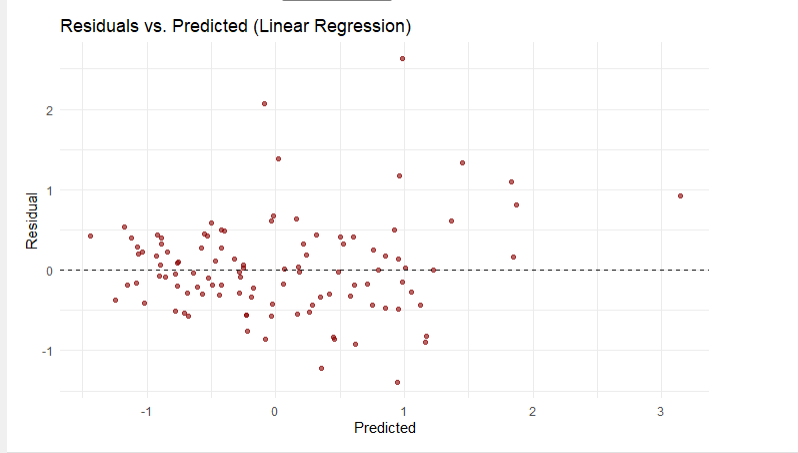
Expecting that high-tech transportation companies would be raiding the ranks of New York’s taxi drivers for drivers. Observing the parking goes from a last of the top-five in the single tree forest to a solid fourth place in the forest. This increase shows that by looking at many trees, the forest is able to find subtle patterns. It contributes to the thesis that parking is a non-negligible amenity premium in this market.

Model Bias & Ensemble Stability

All in all, the tree has a stronger distinguishing signal put into a more “peaked” form: It puts 40+ % of its weight on area, 20% on bathrooms, while the forest is spreading the importance more evenly(~32%,~18%,~9% x5, ~8% x5). This contrast exemplifies the classic bias–variance trade-off: one tree will easily latch onto a few strong predictors (biasing its view of importance), while the forest ensemble will iron out those biases, leaving a more conservative but more robust ranking. When doing feature engineering, it means we should have more faith in the forest ordering to decide on which variables to pay more attention to or transform more.







Actual vs. Predicted Values

Although the actual vs. predicted plot indicates that the models fits the data well overall (most points falling near the diagonal), closer examination reveals a systematic bias:

* **High-End Underprediction:** When high-priced houses are predicted to be lower priced-(basically regression to the mean), perhaps because of the lack of features in the data available to the model in luxury-tier properties or some nonlinear function of sizes.
* **Low-End Overprediction**: Properties with low cost are underestimated, which seems to reflect a floor effect, or suboptimal representation of low-end restrictions in the data.
* **Mid-Range Scatter**: Higher scatter in the middle of the market (between –0.5 and +2 standard units) indicates higher model uncertainty, possibly because of unmodelled variation such as renovation quality, micro-location, or architectural style.

Model Limitations and Paths Forward

Although the linear model is an effective and transparent model, some of its drawbacks are centeredness-bias and inability to describe extreme values, for which one may consider the following improvements:

* **Nonlinear Terms**: Using quadratic/spline transformations for area or bathrooms resolves the shape of the response curve.
* **Tiered Modeling**: Split the data into low/mid/high priced homes and estimate different models (or including interaction terms)
* **Feature Enrichment**: Incorporate geographic, school quality, or architectural characteristics that probably explain variation not captured in core variables.

Residual Analysis Confirms Homoscedasticity with Pockets of Concern

The residuals vs. fitted plot suggests a generally homoscedastic model (errors are spread equally across the value 0), but it indicates two problems:

* **Mild Underprediction for Mid-Range Properties**: Residuals centrally are slanted a bit towards underprediction for predicted values ranging between –0.5 and +0.5, especially for the systemically undervalued “average” houses— due perhaps to lacking some middle-of-the-market features (or maybe location quirks).
* **Extreme Underprediction for Certain Homes**: We have some observations with residuals > +1.5, which could be extreme cases where the model doesn't capture premium value for whatever reason (missing features, such as school district, view, or year built).
* **Takeaway**: Though the residuals look sensible overall, targeted feature engineering (probably based on neighborhood or home age) may fix some of these larger misses and rein in mid-tier predictions.

###### Conclusion and Recommendations

Key Conclusions

Area, bathrooms, and stories are main drivers of price in terms of structure, but they are not proportional one to the other as they have a nonlinear effect on pricing. For instance, the price increases dramatically for space within the mid-market and scales back out for very large sizes. Likewise, the costanness of having more than two additional bathrooms flatlines.

There were trade-offs, according to model comparison:

* **Random Forest** was able to capture nuanced relationships and not overfit any individual feature.
* **Linear Regression** was more interpretable, but it did not predict as well at the outer bounds, and it often overfit less-important variables (e.g., bedrooms, guestroom).
* **Partial Dependence Plots** and residual analyses revealed threshold effects and biases in all price segments but particularly in the middle to high end of the market.

Practical Recommendations

**For Real Estate Developers and Sellers:**

Concentrating on reaching certain amenity benchmarks (for example 2 bathrooms 2 parking spots air conditioning). These have obvious cost advantages. Don’t throw money into features with diminishing returns, like additional bathrooms or bedrooms when the market only pays for so many at a home of its size.

Think vertical growth — extra stories may boost value, especially in denser or higher-end areas.

**For Modelers & Analysts:**

Supplement linear models with nonlinear components, such as splines or polynomial terms, or give the segmented price tier modeling a try to cope with the bounds more effectively.

Additive context information like age of the property, neighborhood score, or renovation status to increase mid-range accuracy and to minimize residual variance.

**For Policy Makers and City Planners:**

Observations corroborate the price effect of location and facilities (e.g., main road access) that justifies investment in the quality of neighbourhood and public transport provision.

Amenity gaps — whether the absence of climate control or off-street parking — may be a sign of inequality in terms of the quality of housing that could also inform affordable housing efforts.

While “larger homes cost more” is a consistent theme, the real story is about feature combinations, market expectations, and non-linear relationships. By mixing model performance with economic intuition and visual diagnostics, it created not only a powerful predictive framework, but also a blueprint for smarter valuation, investment, and feature targeting in the U.S. housing market.

#### 