Evaluating Factors that Drive Boston Residential Property Value

By: Ruben Berganza

Understanding the factors that go into the valuation of properties in Boston is important for homebuyers to be able to assess fair prices, for sellers to be able to maximize their returns, and for city officials that are making fiscal policies that affect housing. With a median residential property valued at $728,400 in Boston in 2025, even small mistakes in predicting housing value can have significant financial consequences. My analysis addresses a fundamental question: Can we predict Boston housing prices based on readily available property characteristics like size, location, and physical features? Using the City of Boston Assessing Department's property assessment data for each year from FY2023 to FY2025, I built and compared five linear regression models to determine which factors most strongly predict property values. The analysis includes 136,511 residential properties from 2025 for single-year predictions and 405,841 properties across all three years for examining price trends over time. The findings of my analysis challenge some traditional and common assumptions related to real estate. While "location, location, location" is often noted as the most important factor of property value, my analysis reveals that physical characteristics, particularly living area and number of rooms, are far more predictive of value than neighborhood or ZIP code. Size-based features alone explained 37% of price variation, while location features explained only 5%.

Data and Methods

Data Collection

I downloaded property assessment data from the City of Boston Assessing Department through data.boston.gov. The datasets cover three fiscal years: FY2023 (180,627 properties), FY2024 (182,242 properties), and FY2025 (183,445 properties). Each dataset contains necessary property information including assessed values, physical characteristics, location details, and property type classifications.

Data Cleaning and Filering

I had to clean a lot of the data before analyzing it. Properties with $0 assessed values (churches, government buildings, charitable organizations) were removed because they don't reflect market conditions. I also filtered to residential properties only to focus exclusively on residential property types like single-family homes (R1), two-family homes (R2), three-family homes (R3), four or more family homes (R4), and residential condominiums (CD). Commercial, industrial, and mixed-use properties were excluded. This resulted in 136,511 residential properties for 2025 and 405,841 properties across all three years. To handle missing values in the key features, I filled them with median values for continuous variables, like living area and bedrooms, or mode values for categorical variables like overall condition. The location features (ZIP code, neighborhood) and property type were also converted to numerical values using the factorization method in order to include them in the linear regression models. As for cleaning the target variable, "TOTAL_VALUE" contained commas in the numbers (e.g., written as "799,000"), which I had to looked up how to handle this and used the .str.replace() method to remove commas before converting it to numerical values.

Variables and Modeling

The target variable for all models was "TOTAL_VALUE", which represents the City of Boston's assessed residential property value in dollars. I grouped the predictor variables into the following three categories:

- Size Features (5 variables): "LIVING_AREA" (square footage), "BED_RMS" (bedrooms), "FULL_BTH" (full bathrooms), "HLF_BTH" (half bathrooms), and "TT_RMS" (total rooms)
- Location Features (2 variables): ZIP_CODE and CITY (neighborhood)
- Property Characteristics (3 variables): YR_BUILT (year built), LU (property type), NUM_PARKING (parking spaces)

I built five linear regression models to analyze the predictive power (R^2) of these different feature categories:

1. Size Features Only (5 features)
2. Location Features Only (2 features)
3. Property Features Only (3 features)
4. Combined Model - 2025 data (all 10 features)
5. Time Series Model - 2023-2025 data (all 10 features across all years)

For each model, I used an 80/20 train-test split and evaluated the actual predictive performance of the model by looking at each model's R^2 score.

Results

Model Performance Comparison

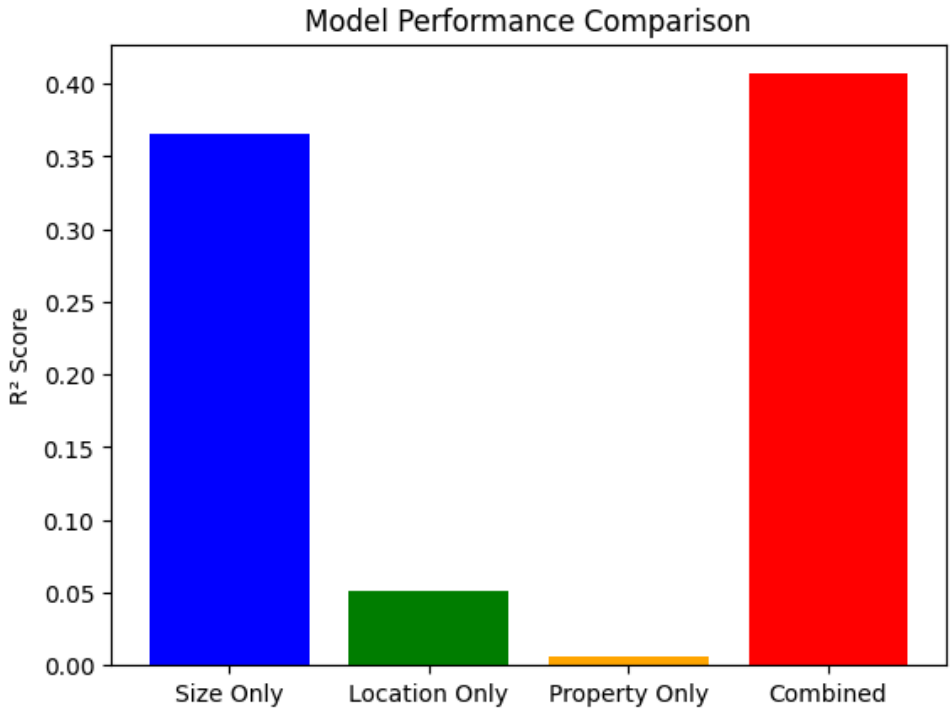
I compared the five linear regression models to determine which property characteristics best predict residential values in Boston. The models varied significantly in predictive power, with R^2 scores ranging from 0.006 to 0.4068.

Size Features Only Model: Using only five physical characteristics, being living area, bedrooms, bathrooms, and total rooms, the model had an R^2 score of 0.365. This means that size features alone explain 36.5% of the variation in property values. This was by far the strongest single category of features.

Location Features Only Model: Using only ZIP code and neighborhood, the model achieved an R^2 score of only 0.051, explaining just 5% of variation. This result directly contradicts the common assumption that "location is everything."

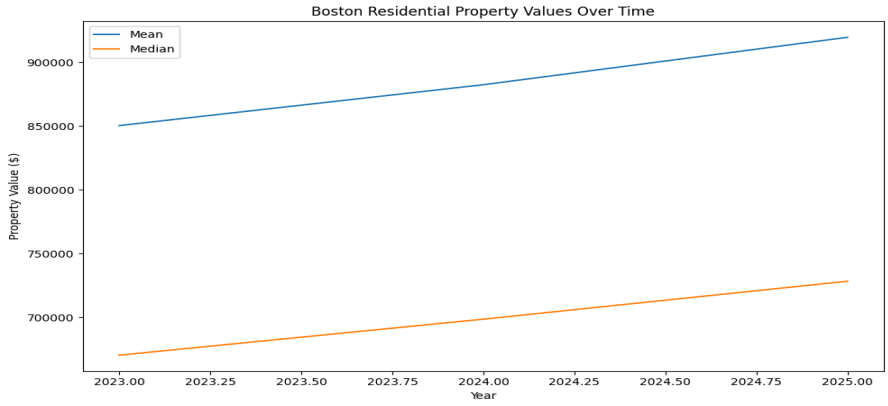
Property Characteristics Model: Using just the year the property was built, property type, and number of parking spaces, this was the worst performing model with an R^2 of 0.006, which is less than 1% and essentially not predictive at all.

Combined All Features Model: Using all 10 features together had the best performance, achieving the highest accuracy at $R^2 = 0.4068$. This means that all of these features combined explain 40.7% of the variation in property values. While this is our best performing model, it unfortunately still cannot explain nearly 60% of the variation in price value.



Time Series Analysis: 2023-2025 Price Growth

Analyzing 405,841 residential properties across three years showed consistent price growth in Boston's housing market. The mean property value increased from $850,308 in 2023 to $919,439 in 2025, an 8.13% increase over two years (so around 4% each year). The median property value showed similar growth, rising from $670,500 to $728,400, an 8.64% increase. This steady growth pattern suggests a relatively stable, but still increasingly pricy, housing market without any dramatic shifts.

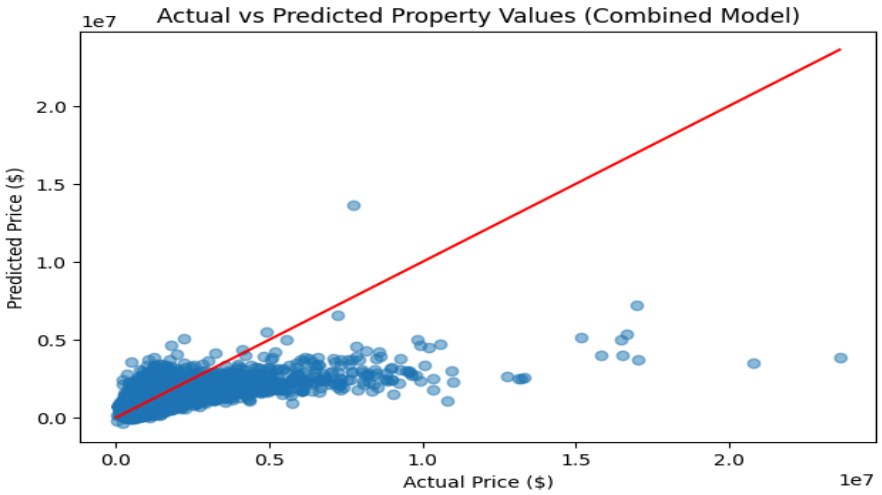


Impact of Adding Time as a Feature

To test whether knowing the year improves predictions, I built a Time Series Model that included “YEAR” as an 11th feature on top of the existing 10 combined features. This model achieved an R^2 of 0.408, only 0.0012 better than the Combined Model without the year variable. This minimal improvement means that the characteristics of the property matter a lot more than market timing for predicting values. In other words, a 1,500 square foot home in Dorchester will be similarly valued relative to other properties regardless of whether it's assessed in 2023 or 2025, even though overall market prices rose during those years.

Model Predictions vs. Actual Values

The scatter plot below shows predictions from the Combined Model, the best performing model, against actual property values. Points falling on the red line would indicate perfect predictions. The plot reveals that the model performs decently well for properties in the $200,000-$500,000 range (where most properties are clustered) but start to become increasingly less accurate with higher valued properties above $500,000, gradually undervaluing the price of the property more as the actual price goes up.



Discussion

Limitations of the Data

My analysis uses assessed property values from the City of Boston, not the actual sale prices, which are typically lower than market prices. This means the model predicts what the city thinks properties are worth for tax purposes, not necessarily what buyers would pay.

The dataset also lacks several factors known to influence housing prices such as quality of schools, crime rates, proximity to public transit, walkability, recent renovations, and other amenities in the neighborhood. The surprisingly low predictive power of the location features (with an R^2 of 5%) might partly reflect this as ZIP code alone cannot really account for important differences within a neighborhood.

Limitations of the Model

The Combined Model's R^2 of 0.407 means it explains around only 41% of price variation, leaving the remaining 59% unexplained. Linear regression assumes that the relationships between features and price are linear, but housing prices may have more complex, non-linear relationships that I am not able to demonstrate with my simpler model.

The model also struggles with high-value properties above $500,000, as it begins to increasingly underpredict their prices the higher it becomes. This might mean that more expensive properties are valued according to slightly different factors, maybe like luxury, unique features, or views, that I cannot demonstrate with the assessment data.

Surprising Findings

The most surprising result was how poorly location predicted property values. The popular sentiment toward housing prices is that "location is everything," so we would assume it would explain much of the price variation. However, the location features explained only 5% of variance while size features explained 36.5%, which is significantly higher. This suggests either of the following possibilities: (1) Within Boston, neighborhoods have relatively similar demand once you account for property size, or (2) simply going by ZIP code and neighborhood name fails to account for what actually makes a location valuable. As a Boston resident, I believe the latter to be the most likely explanation. My home neighborhood of Dorchester is generally considered to be a somewhat undesirable town due to crime, but its large size, proximity to the main city, and recent developments in the northern area of the town have caused its property values there to increase significantly, despite the rest of the town staying relatively the same.

Another surprise was how little adding the 'YEAR' feature improved the model (only 0.1% increase in the R^2). Despite 8% market-wide price growth from 2023-2025, knowing the year barely helped predict individual property values. This might mean that relative property values within Boston remain stable even as the overall market rises.

Conclusion

My analysis challenges the traditional idea that location is the primary driver of housing prices. Using three years of Boston property assessment data, I found that physical size characteristics are seven times more predictive of property values than location features. A combined model achieved 41% accuracy in predicting assessed values.

For people looking to buy a home, this suggests that focusing on physical property characteristics may be more important than fixating on specific neighborhoods when evaluating its value. For policymakers, these findings validate that physical characteristics provide a reasonable basis for tax assessments, though the 59% of unexplained variance indicates other factors like school quality, amenities, and property condition, are also very important but were not shown in the basic assessment data.

The Boston housing market showed steady growth from 2023-2025 with an average annual growth of around 4%, and no drastic shifts. Predictions could be improved by accounting for school ratings, crime, public transit, and actual sale prices, or by using better predictive techniques.