Capturing the True Value of Real Estate in Seattle Through Property Characteristics: A Machine Learning Approach*

Location and Age Prove Key Determinants in Calculating Intrinsic Value

Sameeck Bhatia

December 3, 2024

This paper develops a transparent property valuation model for Seattle's real estate market, using Redfin data and machine learning techniques to estimate intrinsic property values. Results show higher prediction errors for high-value properties and the importance of property age as well as amenities in determining valuation. Additionally, spatial analysis captures the influence of environmental factors on property values, reflecting income and infrastructure differences. These findings help buyers and sellers understand what drives property prices, supporting informed decision-making to navigate a competitive market.

Table of contents

1 Introduction 2

2 Data 3

2.1 Overview 3

2.2 Summary Statistics 4

2.3 Measurement 5

2.4 Outcome Variables 5

2.5 Predictor Variables 6

3 Model 7

3.1 Overview 7

3.2 Setup 8

*Code and data are available at https://github.com/SameeckBhatia/Seattle-Real-Estate

3.3 Limitations 8
3.4 Justification 9
3.5 Interpretation 9
3.6 Validation 10
3.7 Alternate Models 11

4 Results 11
4.1 Higher Prices Associated with Higher Prediction Errors 11
4.2 Similar Distributions to Market Price from the Model 12
4.3 Influence of the Time Effect on Property Valuation 13
4.4 Noticeable Clusters for High-End Properties 14

5 Discussion 15
5.1 Examining Racial Bias in Real Estate Price Estimates 15
5.2 Using Alternative Algorithms for Nonlinear Price Trends 15
5.3 Including External Factors into Property Valuation Models 16

Appendix 17

A Additional Data Details 17
A.1 Raw Data Dictionary 17

B Additional Model Details 18
B.1 Model Diagnostics 18
B.2 Alternate Model Summaries 19

C Surveys, Sampling, and Observational Data 20

References 22

1 Introduction

Seattle's real estate market has experienced significant fluctuations since the start of the millennium. Despite these booms and busts, property prices in Seattle remain among the highest in the United States (Jones 2024). Demand has steadily increased, and even during periods of lower demand, the market continues to attract buyers. At current price levels, more potential homebuyers are seeking reliable information to find the best deals. With advancements in technology and the increasing availability of data, tools for assessing property values have become essential. One important tool is an estimate of a property's fair value, which can help buyers determine whether they are overpaying or underpaying relative to the market.

This paper focuses on developing an accurate and accessible property valuation model tailored to Seattle's housing market. The model incorporates current market data to provide fair value estimates for individual properties. The goal is to inform readers about the key factors driving real estate prices in Seattle while empowering them with a tool to make better home-buying decisions. By applying machine learning and statistical techniques, the model aims to ensure accuracy and reflect the intrinsic value of properties rather than merely their market prices. Existing valuation tools, such as Zillow's "Zestimate" (Zillow 2024) and Redfin's proprietary estimates, have limitations. These models are often closed-source, leaving potential users unable to access the methodology or even the results without cost. This creates a significant gap in the availability of open, free, and transparent valuation tools. The absence of such tools denies buyers the advantage of comprehensive, unbiased information to assess property values independently. To address this gap, data was collected from Redfin's semi-public dataset on current Seattle property listings. The analysis was conducted using the R programming language for data cleaning, testing, model creation, and result interpretation.

The primary estimand of this paper is the intrinsic value of a property. This value is determined using a linear regression model applied to observational data collected across Seattle. The analysis aims to identify the true value of a property that a buyer should consider paying and a seller should consider accepting. Additionally, it examines the main factors influencing a property's intrinsic value and allows for comparison to market values.

Preliminary findings indicate that the model effectively predicts property prices, though deviations occur at extreme price ranges. These discrepancies suggest that additional factors, such as high-end amenities and neighborhood characteristics, play a significant role in influencing property value. Understanding these dynamics is important in incorporating socio-economic and structural variables into property valuation models, ultimately improving their accuracy and real-world applicability. The paper is structured as follows: Section 2 introduces the dataset and variables. Section 3 outlines the model design and its significance. Section 4 presents the model's predictions and actionable insights for buyers. Finally, Section 5 explores potential applications of the model in other cities, factors influencing property prices, and the generalizability of the model.

2 Data

2.1 Overview

The data used in this paper comes from Redfin, a real estate brokerage and mortgage company, and represents current property prices across Seattle. While the data was collected from Redfin (2024), similar data could have been sourced from competitors like Zillow and Realtor.com, with Zillow being the largest of the three. If data from Zillow or Realtor.com had been used, the listings might vary slightly since each platform likely features different sellers and listings. However, Redfin was chosen as the source because it is the only brokerage firm that allows public downloads of its listings, helping to avoid violations of data extraction policies.

The dataset includes only properties within Seattle's official city boundaries and covers condos, townhouses, and single-family homes currently on the market. The raw data contains 27 variables, 18 of which have been included in the cleaned dataset. These include variables such as MLS number, number of bedrooms, neighborhood, and geographic coordinates. Additionally, some variables were constructed specifically for this analysis, such as `half_bath`, `property_age`, and `price_sqft`. The first two were added to enhance the dataset and valuation model, while `price_sqft` was created as an alternate response variable. Further details about these variables can be found in Appendix A. The data was analyzed using R (R Core Team 2023) and the tidyverse (Wickham et al. 2019) package, while visualizations have been created using tidyverse.

2.2 Summary Statistics

Table 1

Statistic	N	Mean	St. Dev.	Min	Median	Max
beds	1,361	2.69	1.42	0	3	12
baths	1,361	2.03	1.02	0	2	13
sqft	1,361	1,706.78	1,192.46	223	1,373	13,710
year_built	1,361	1,987.55	37.51	1,890	2,002	2,024
days_on_market	1,361	64.31	65.10	1	49	878
hoa_month	1,361	323.15	623.58	0	13	10,281
price	1,361	1,107,431.00	1,575,578.00	199,000	775,000	39,950,000
half_bath	1,361	0.34	0.47	0	0	1
property_age	1,361	36.45	37.51	0	22	134
price_sqft	1,361	631.41	319.11	205.09	590.71	8,567.45

Table 1 presents summary statistics for all original and derived numeric variables in the dataset. The mean number of bedrooms is 2.69, with a median of 3, while the mean number of bathrooms is 2.03, with a median of 2. This indicates that the typical listing has around three bedrooms and two bathrooms, commonly seen in townhouses and single-family homes. The average property size is approximately 1,710 square feet, with a median of 1,373 square feet, suggesting a positive skew in property size due to a few larger homes in the dataset. The average time a property remains on the market is around 64 days (just over 2 months), with a maximum of 878 days (nearly 2.5 years), indicating low demand in Seattle's real estate market, especially since the data contains only active listings. The mean and median property prices are $1,107,431 and $775,000, respectively, while the mean and median price per square foot are $631.41 and $590.71, respectively. These figures highlight Seattle as one of the most expensive residential markets in the United States.

2.3 Measurement

In the United States, buyers and sellers have the freedom to select the real estate agent or brokerage firm they wish to work with for transactions. Consequently, agents often represent multiple listings within their region or city. Agents receive detailed information on each property from real estate appraisers, who measure variables such as the number of bedrooms, bathrooms, and square footage (National Association of Realtors 2024). These measurement practices, except for price, are strictly regulated to ensure accuracy for all stakeholders.

For property prices, appraisers typically estimate values based on the prices of recently sold comparable properties and the specific characteristics of the property. This valuation process is less regulated, as it serves primarily as a reference point for buyers and sellers. Real estate agents may gather price estimates from multiple appraisers to calculate an average. The prices observed in the data, although guided by these values, are ultimately set by the seller. All this information is uploaded to the Multiple Listing Service (MLS), a private database accessible only to agents and brokerage firms via subscription fees (Bankrate 2024). However, U.S. laws allow companies like Redfin, Trulia, and Zillow to extract and share MLS data with the public, fostering competition and transparency.

2.4 Outcome Variables

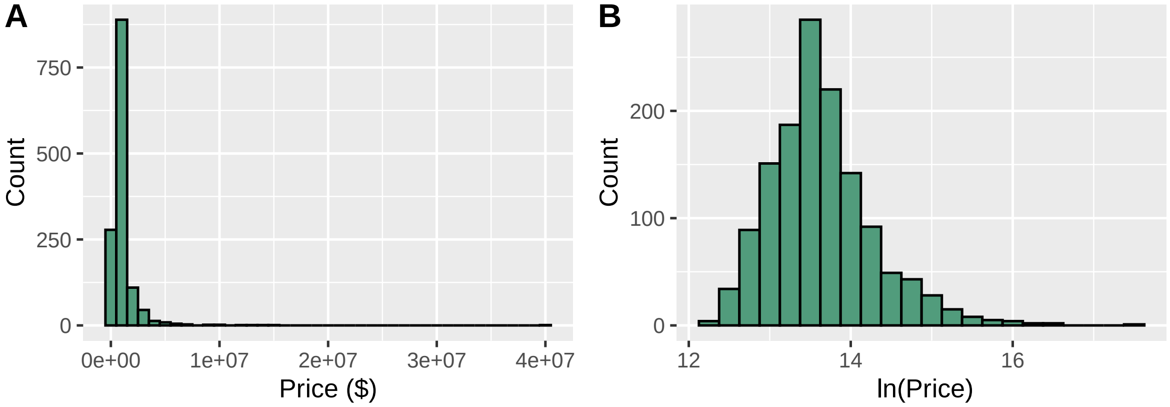

Figure 1: Heavily Skewed Property Prices and Skewed Still for Logarithmic Prices

The primary goal of this paper is to estimate the outcome variable, price, using a valuation model. This variable represents the market value of active listings at the time of data collection, focusing exclusively on properties located in Seattle. Figure 1 provides insights into the distribution of property prices. Plot A shows that property prices are highly skewed, with a maximum value near $40,000,000 and several listings exceeding $5,000,000. To better understand the distribution, Plot B presents the logarithmic transformation of property prices.

While the transformed distribution is less extreme, it remains skewed, indicating the presence of properties with exceptionally high valuations in the Seattle real estate market.

2.5 Predictor Variables

Property Type (`property_type`): This categorical variable identifies the type of property for each observation, based on classifications set by the MLS. The three main property types included in the analysis are “Condo/Co-op,” “Single Family,” and “Townhouse.” These classifications are important for valuation, as different property types possess distinct features that influence their market value.

Number of Bedrooms (`beds`): This numeric variable represents the count of full bedrooms in a property, as measured by appraisers. It is a key factor in the analysis, as more bedrooms often correlate with greater living space and the potential to accommodate larger households.

Number of Bathrooms (`baths`): This variable represents the total count of full bathrooms in a property, as measured by appraisers. Properties with more bathrooms typically appeal to larger households, and is sometimes caused by the number of bedrooms in a property, leading to correlated prices.

Property Size (`sqft`): Measured in square feet, this variable reflects the total size of the property and is determined by appraisers. As the United States predominantly uses the imperial system, square footage is a standard unit. Larger properties generally hold more value, making this a significant predictor in property valuation.

Property Age (`property_age`): This derived variable calculates the age of a property in years, based on the difference between the year of data collection and the `year_built`. Age is a vital consideration in determining price, as newer properties are often smaller due to rising construction costs and increasingly strict zoning regulations.

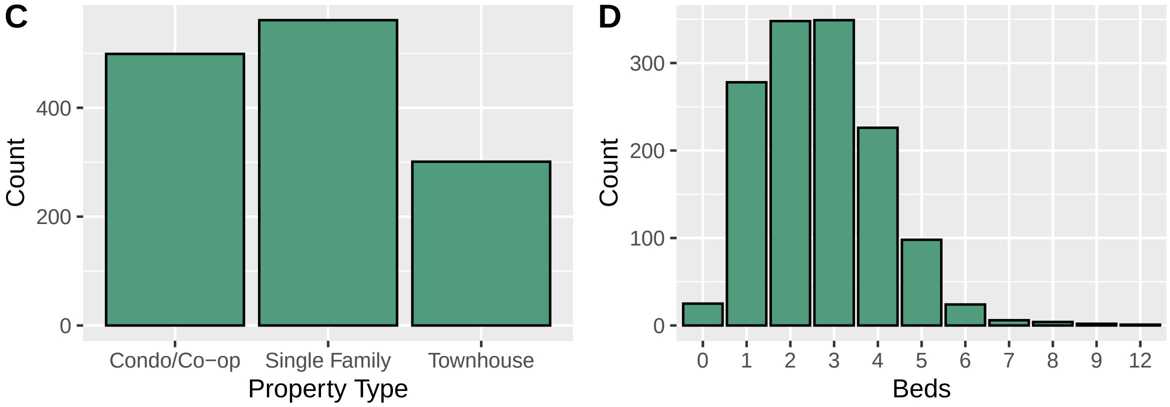

Figure 2: Single Family Homes and 2 or 3 Bedrooms Properties Most Listing on the Market

Figure 2 represents the distributions of property types and number of bedrooms. Plot C shows that single-family homes are the most common, followed closely by condos, with townhouses being the least prevalent. This likely reflects the mix of apartments and houses present in Seattle's real estate market. Plot D highlights that properties with 2 or 3 bedrooms are most frequent, while some properties even have north of 7 bedrooms. This is likely because condos typically feature 1-2 bedrooms, whereas single-family homes and townhouses commonly have 3-4 bedrooms.

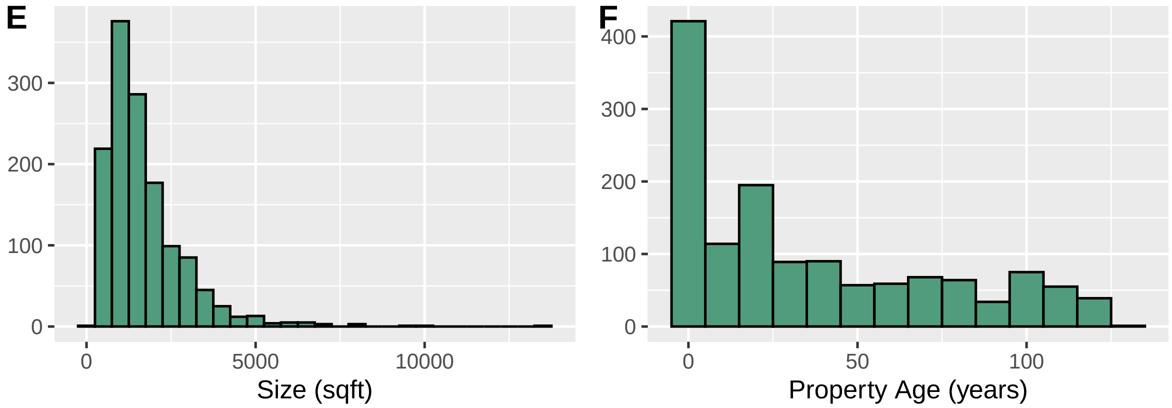

Figure 3: Skewed Distribution of Sizes and Most Properties Built Over the Last Decade

Figure 3 displays the distributions of property size and age. Plot E shows that property size is positively skewed, with most properties under 2,500 square feet and a peak between 700 and 1,200 square feet. Properties exceeding 5,000 square feet likely represent luxury homes or mansions. Plot F indicates that property age is also positively skewed, with the most common properties being less than 10 years old. The second most common group is 20-30 years old, likely reflecting construction surges before the 2008 financial crisis and real estate market downturn.

3 Model

3.1 Overview

The model uses a multiple linear regression approach to estimate property values in Seattle. It incorporates features like the number of bedrooms (`beds`), square footage (`sqft`), homeowner association fees (`hoa_month`), and an interaction term (`sqft` × `beds`) to predict prices. These characteristics were selected for their significance in influencing property valuations. The final model was chosen after thorough validation, achieving a good balance of accuracy and interpretability.

3.2 Setup

The linear regression model can be represented as:

$$\text{price} = \beta_0 + \beta_1 \cdot \text{beds} + \beta_2 \cdot \text{baths} + \beta_3 \cdot \text{sqft} + \beta_4 \cdot \text{hoa\_month} + \beta_5 \cdot (\text{sqft} \cdot \text{beds})$$

, where the coefficients are described as:

β_0: Intercept

The predicted value of `price` when all predictors (beds, sqft, hoa_month, sqft · beds) are zero.

β_1: Coefficient for `beds`

The effect of adding one more bedroom on price, holding other factors constant (except for the interaction term).

β_2: Coefficient for `baths`

The effect of adding one more bathroom on price, holding other factors constant.

β_3: Coefficient for `sqft`

The effect of increasing square footage by one unit on price, holding other factors constant (except for the interaction term).

β_4: Coefficient for `hoa_month`

The effect of a one-unit increase in monthly homeowner's association fees on price, holding other factors constant.

β_5: Interaction between `sqft` and `beds`

Represents how the relationship between sqft and `price` changes depending on the number of bedrooms.

3.3 Limitations

This model has a few limitations that should be considered when interpreting its predictions. First, it is trained on cross-sectional data rather than longitudinal data. As a result, the model is designed to provide accurate valuations for properties in the near future (typically less than a year), assuming minimal price fluctuations. Significant market changes over time would reduce the model's accuracy, requiring frequent updates with new cross-sectional data to maintain reliability. Additionally, the model is specifically tailored to Seattle's real estate market. While it might perform adequately in nearby cities with similar market characteristics, it is unlikely to generalize well to cities in other states, such as Los Angeles or New York, due to differing property attributes and pricing dynamics in those regions.

3.4 Justification

The four features included in the model are `beds`, `sqft`, `hoa_month`, and `sqft:beds`, as they were the most significant in influencing the model's property valuations. The beds feature was selected because the number of bedrooms is a key determinant of a property's utility and appeal to buyers, directly impacting its market value. The sqft feature, representing the total square footage, is a fundamental metric for assessing a property's size and, consequently, its worth. The hoa_month feature accounts for monthly homeowner association fees, which can significantly affect the affordability and desirability of properties, particularly in condominiums or communities with shared amenities. Finally, the interaction term sqft:beds captures the relationship between the size of the property and the number of bedrooms, highlighting how the distribution of space impacts valuation.

Figure 4 represents the causal relationships between the variables analyzed in the regression model. It illustrates that property price is influenced by the number of bedrooms, bathrooms, square footage, and HOA fees. Square footage itself depends on the number of bedrooms and bathrooms, while HOA fees are linked to square footage. The directed arrows capture the assumed causal pathways, emphasizing that the effects of bedrooms and bathrooms on price are partly mediated through square footage.

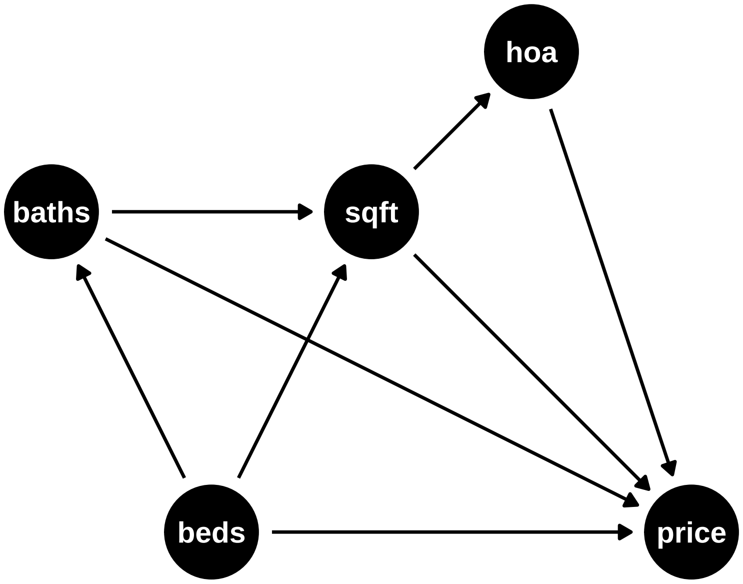

Figure 4: Multiple Causal Relationships Among Model Variables

3.5 Interpretation

Table 2 shows the values of the model's coefficients mentioned in the setup. The intercept of −66,054.23 suggests that, without any additional predictors, property prices would begin

Table 2

	Final Model
(Intercept)	−66 054.23
	(34 398.40)
beds	55 091.77
	(13 863.34)
baths	41 740.69
	(20 967.02)
sqft	592.23
	(22.87)
hoa_month	66.39
	(17.85)
beds × sqft	−38.66
	(3.33)
Num.Obs.	952
R2	0.717
R2 Adj.	0.716
RMSE	331 563.51

at this value, though it has limited standalone interpretation. The coefficient for `beds` is 55,091.77, meaning that, holding other factors constant, each additional bedroom is associated with an addition in price. However, this effect is modified by the interaction term `beds` × `sqft`, which has a negative coefficient of −38.66, indicating that larger properties have diminishing prices. The coefficient for `baths` is 41,740.69, meaning that for each additional bathroom, the predicted value of the price increases by around $41,740.69, holding all other variables constant. This suggests that bathrooms have a strong positive impact on the value of the property. The `sqft` variable has a strong positive impact (592.23 per additional square foot), highlighting size as a key driver of value. The `hoa_month` coefficient (66.39) shows a modest positive association with price, possibly reflecting higher costs in premium communities. With an adjusted R^2 of 0.716 and an RMSE of 331,563.51, the model has a decent fit.

3.6 Validation

The linear model was created using R (R Core Team 2023) to fit the data and generate predictions, while the MLMetrics package (Yan 2024) was used to evaluate performance. A train-test split was created with the rsample package (Frick et al. 2024), with the model trained on the 70% of the data and validated using out-of-sample testing on the remaining 30%. Key evaluation metrics included the Root Mean Square Error (RMSE) and R^2 The model achieved an RMSE of 331,564, indicating the average prediction error in dollar terms, and R^2

of approximately 0.717, reflecting the proportion of variance in property prices explained by the model. Further diagnostics and information are provided in Appendix B.

3.7 Alternate Models

Several alternative models were considered before selecting the final one. The first, a "full model," included all variables in the cleaned dataset, achieving an RMSE of 333,693.77 and an R^2 score of 0.714. While this model used all available information, some variables were statistically insignificant. A reduced model was then tested by retaining only significant variables, resulting in an RMSE of 333,736.71 and an R^2 score of 0.713. Though concise, it performed similarly to the full model. The final model, incorporating an interaction term, outperformed both alternatives with an RMSE of 331 563.51 and an R^2 of 0.717.

4 Results

4.1 Higher Prices Associated with Higher Prediction Errors

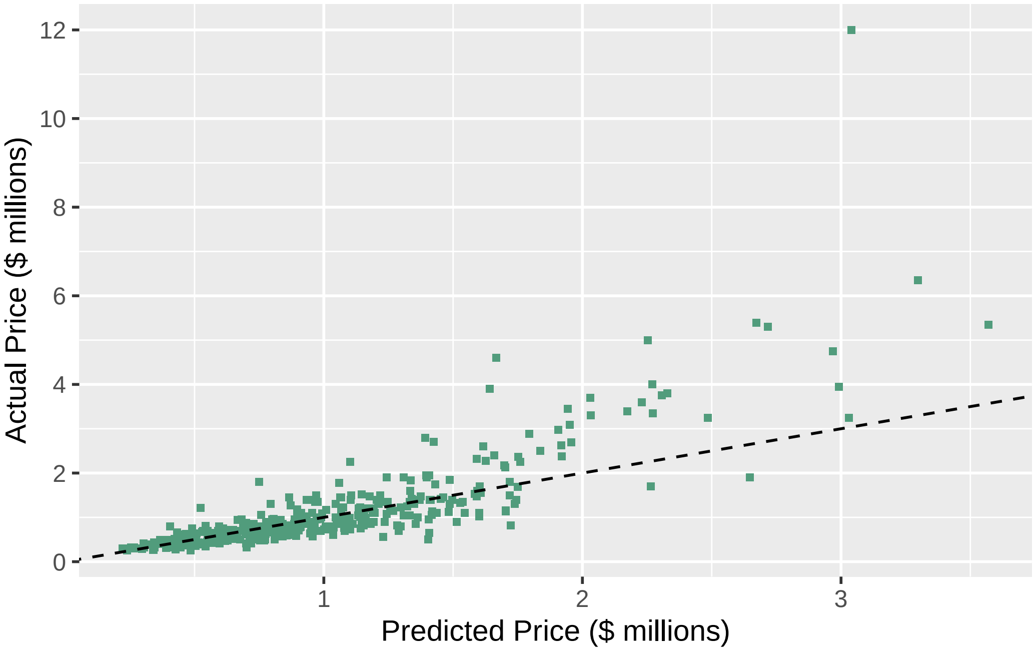

Figure 5: Large Uncertainty in the Predicted Values of Luxury Homes

Figure 5 shows the model-predicted prices alongside the actual price observations from the data. The dashed line represents the line of equality, where the predicted price matches the actual property price. Some data points deviate from this line, particularly at lower values, which is expected since no model can perfectly capture all variations in the data. In this case, the deviations may result from missing factors that could enhance the property valuation model. However, when property values exceed approximately 2 million dollars, the actual prices deviate significantly from the line of equality. This indicates higher prediction errors, likely due to a combination of internal and external factors. Internal factors might include additional amenities (e.g., pools, large gardens, or extra rooms) or premium furnishings (e.g., expensive flooring or tiling), which increase lot size or property desirability (Realtors 2024), driving up prices. External factors may include the property's location in high-income neighborhoods or proximity to desirable features such as beaches, shopping malls, or tourist attractions. These influences are analyzed further in Figure 8.

4.2 Similar Distributions to Market Price from the Model

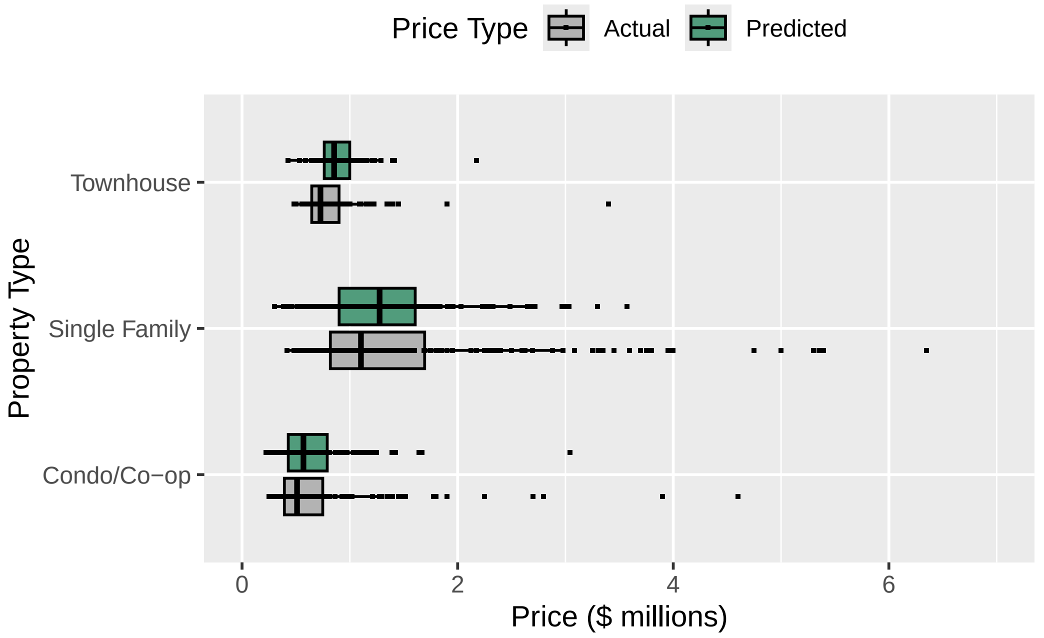

Figure 6: Single Family Homes Valued the Highest Among All Property Types

Figure 6 illustrates the distributions of actual and predicted prices based on MLS-classified property types. Condos have the lowest median prices, followed by townhouses, with single-family homes being the most expensive. This pricing trend is primarily influenced by the size of each property type, which plays a significant role in determining value (List 2024). The predicted and actual distributions for all property types show similar shapes, although there are differences in their centers and outliers. The model predicts slightly higher intrinsic values for townhouses and condos compared to the actual prices, whereas there is a notable difference in the median price for single-family homes. This discrepancy is likely due to the real estate market being in a cooling phase (Estate 2024), with prices remaining relatively stable over the past six months. Additionally, while market prices are subject to short-term fluctuations, intrinsic values tend to remain more consistent.

4.3 Influence of the Time Effect on Property Valuation

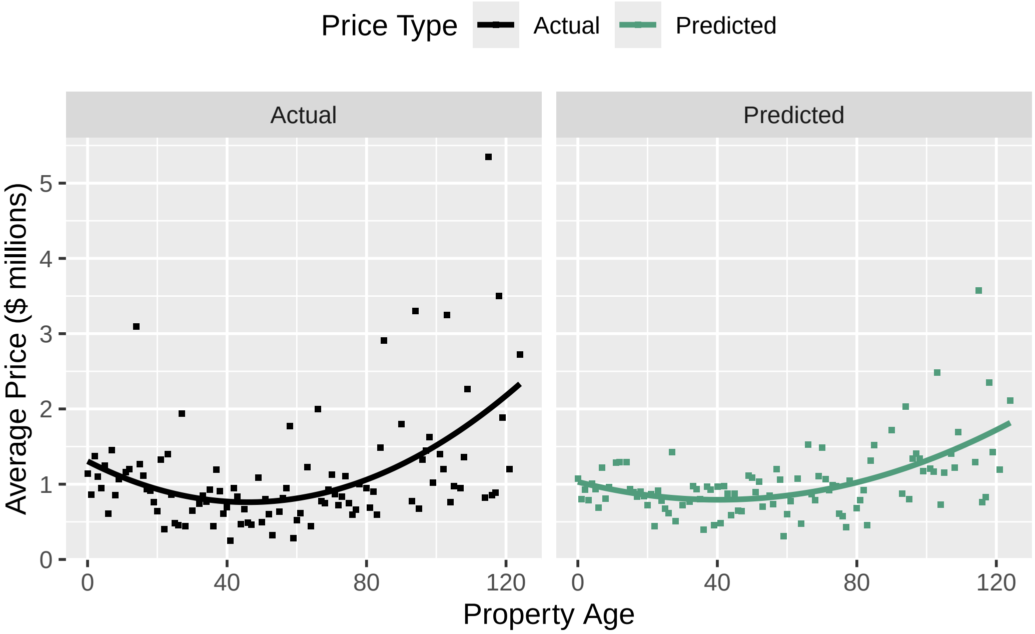

Figure 7: 40 Year Old Homes Among the Lowest in Actual and Predicted Values

Figure 7 highlights the trend between property age and the predicted and actual average prices. The relationship can be divided into three stages. First, the average price for brand-new properties starts at around 1 million dollars, followed by a gradual decrease as the properties age. This decline is expected, as older properties typically lose value due to higher upkeep costs and reduced desirability (Coulson and McMillen 2008). Next, the average price bottoms out when properties reach around 40 years of age. This likely represents a threshold where property age significantly affects market value (for actual prices) or true value (for predicted prices). In the third stage, average prices increase for properties over 50 years old, continuing to rise for homes built a century ago. While this may seem counterintuitive given the earlier trend, it likely reflects the premium placed on vintage properties. Such homes often have historical significance, unique architectural features, or are built on highly desirable land, all of which enhance their perceived and intrinsic value.

4.4 Noticeable Clusters for High-End Properties

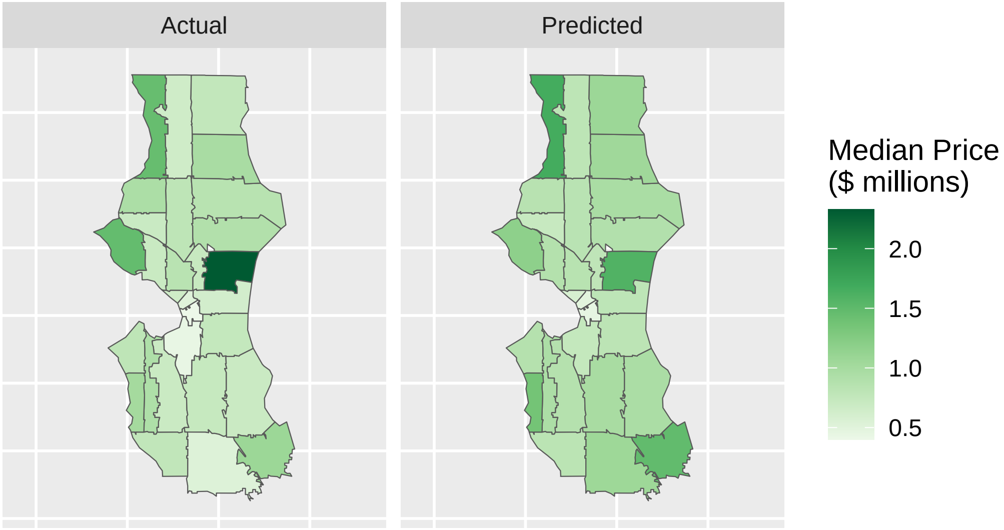

Figure 8: Highest Predicted Listing Prices in Seattle's Largest Gated Communities

Figure 8 presents the spatial distribution of median property values across Seattle, categorized by ZIP code. Two distinct patterns emerge from both maps: first, two ZIP codes stand out with the highest median listing prices; second, property values decrease the farther south they are in the city. The model captures these patterns well, showing similar clusters. The two ZIP codes with the darkest shades correspond to the Broadmoor and Broadview neighborhoods, which are 85- and 1,100-acre gated communities with affluent residents and amenities such as golf courses. These communities share features like higher incomes, better security, and a higher standard of living, enabling them to support significantly higher property values. Conversely, neighborhoods in southern Seattle tend to have lower incomes, higher crime rates, and less development, and evidently lower property values. These factors collectively explain the disparities in property values across the city. These disparities reflect broader socio-economic trends in urban areas, where wealthier neighborhoods often attract more investment in infrastructure, public services, and amenities, driving property values higher (Kim 2021).

5 Discussion

5.1 Examining Racial Bias in Real Estate Price Estimates

The integration of algorithmic estimates like Zillow's Zestimate into real estate markets offers potential to reduce racial biases in property valuation (Yu 2020). Historically, systemic practices such as redlining have resulted in undervaluation of properties in minority neighborhoods, reinforcing economic disparities. The Zestimate, which aggregates data from multiple sources and provides consistent valuations across neighborhoods, appears less influenced by these historical biases. As a public and standardized information source, it helps market participants align on property values, leading to reduced racial disparities in final sale prices compared to initial list prices. This effect is especially meaningful in addressing the "white premium" often observed in traditional valuation methods.

However, while the Zestimate mitigates some biases, it does not entirely eliminate racial inequities in the housing market. Factors like historical undervaluation and socio-economic disparities remain deeply embedded, requiring more targeted interventions beyond algorithmic tools. Moreover, reliance on algorithmic estimates can sometimes oversimplify the complex factors that influence property values, such as cultural significance or community dynamics, which are more difficult to quantify. Nevertheless, the use of transparent and publicly available valuation tools represents a positive step toward a more equitable real estate market.

5.2 Using Alternative Algorithms for Nonlinear Price Trends

Machine learning has become a fundamental method for modeling non-linear trends in property prices, capturing the complex relationships between property characteristics, location, and market factors. Traditional methods, such as linear price models, often struggle with these complexities, especially when spatial dependencies at regional levels influence property values. Advanced algorithms like Random Forest and Gradient Boosting have shown promise in addressing these challenges. By using data at finer geographical scales, such as statistical sub-areas, these models achieve greater accuracy, particularly when predicting price differences between housing types like apartments and standalone homes. This spatially informed approach helps stakeholders understand localized price trends and refine policy-making (Gao et al. 2022).

Emerging methods like Artificial Neural Networks and Fuzzy Inference Systems further enhance the ability to capture non-linear patterns in real estate data. Recent developments in fuzzy regression techniques have demonstrated superior performance in predicting prices while maintaining lower computational costs compared to older models. These approaches excel at interpreting complex interactions between pricing attributes and property values, offering improved precision even for less structured data like textual property descriptions. Such innovations highlight the growing potential of machine learning to address long-standing challenges in real estate valuation, providing better tools for buyers, sellers, and policymakers to navigate market dynamics (Sarip, Hafez, and Daud 2016).

5.3 Including External Factors into Property Valuation Models

Incorporating external environmental factors into property valuation models offers great advantages, such as a more comprehensive understanding of real estate prices influenced by location-specific qualities (Din, Hoesli, and Bender 2001) (Abelson 1979). Geographic information systems (GIS) provide detailed data on factors like neighborhood quality, proximity to amenities, and environmental conditions, enabling models to quantify their impact on property values. For example, positive attributes such as good views or access to shops can enhance valuation, while disamenities like noise pollution may reduce it. These factors add depth to traditional models, ensuring they reflect localized preferences and market trends, which is particularly helpful for urban planning and decision-making.

However, using external factors also poses challenges. Data collection for environmental parameters is resource-intensive and may require frequent updates to stay relevant. Additionally, interpreting these factors can be complex, as their effects vary by region and buyer priorities. For instance, while some buyers prioritize neighborhood aesthetics, others might value proximity to public transport more. Simplifying these influences into a single measure, like a geo-index, may overlook important complexities, leading to inaccuracies in valuation. Furthermore, reliance on advanced tools like artificial neural networks increases model complexity, requiring specialized expertise and computational power, which may not be feasible for all users.

Appendix

A Additional Data Details

A.1 Raw Data Dictionary

Variable	Description
SALE TYPE	Type of sale (e.g., MLS listing, new construction home, new construction plan).
SOLD DATE	Date when the property was sold.
PROPERTY TYPE	Type of property (e.g., condo, single-family house, townhouse).
ADDRESS	Full address of the property.
CITY	City where the property is located.
STATE OR PROVINCE	State or province where the property is located.
ZIP OR POSTAL CODE	ZIP code of the property location.
PRICE	Sale price of the property.
BEDS	Number of bedrooms in the property.
BATHS	Number of bathrooms in the property.
LOCATION	Neighbourhood of the property.
SQUARE FEET	Total square footage of the property.
LOT SIZE	Lot size in square feet
YEAR BUILT	Year the property was built.
DAYS ON MARKET	Number of days the property has been listed on the market.
$/SQUARE FEET	Price per square foot of the property.
HOA/MONTH	Monthly Homeowners Association (HOA) fee, if applicable.
STATUS	Current status of the property (e.g., sold, pending, active).
NEXT OPEN HOUSE START TIME	Start time of the next scheduled open house, if available.
NEXT OPEN HOUSE END TIME	End time of the next scheduled open house, if available.
URL	URL to additional property details.
SOURCE	Source of the property data (e.g., MLS, Zillow).
MLS#	Multiple Listing Service identification number for the property.
FAVORITE	Indicates whether the property is marked as a favorite (e.g., Y/N).
INTERESTED	Indicates whether the user has expressed interest in the property (e.g., Y/N).
LATITUDE	Latitude of the property location for geospatial analysis.
LONGITUDE	Longitude of the property location for geospatial analysis.

B Additional Model Details

B.1 Model Diagnostics

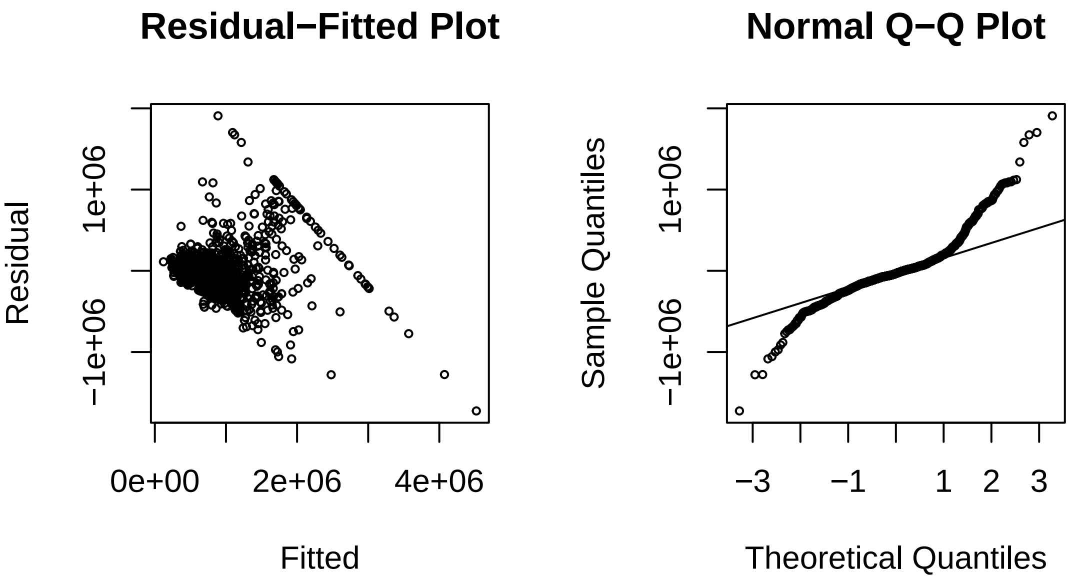

Figure 9: Visible Violations of Constant Variance and Normality Assumptions

Figure 9 presents the residual-fitted and quantile-quantile plots for the final model, used to assess the validity of linear regression assumptions. These diagnostics were conducted to identify any violations of key assumptions. Notably, two primary assumptions were tested, and one key modification was made to the model to improve its fit to the data. Specifically, the model was fitted using the analysis data with imputed price values to the 95^{th} percentile, addressing the presence of significant price outliers in the dataset. This adjustment is evident in the diagonal line appearing to the right in the residual-fitted plot. The imputation process resulted in a substantial increase in the model's R^2, rising from approximately 0.28 to 0.72, as detailed in Section 3. The primary assumption violations observed were related to constant conditional variance and normally distributed errors. These issues are reflected in the left plot, where fanning indicates non-constant variance, and deviations from the diagonal line suggest departures from normality. To address these violations, transformations of the variables, such as using power terms, could be considered; however, this would compromise the model's interpretability. For the final model, interpretability was prioritized, with transformations remaining a topic for future exploration.

Table 4

	Full	Reduced
(Intercept)	276 849.796	268 032.931
	(34 207.507)	(28 306.760)
property_typeSingle Family	198 743.750	198 184.223
	(42 548.641)	(37 982.409)
property_typeTownhouse	−50 299.334	−52 422.284
	(38 201.911)	(36 368.931)
beds	617.997	
	(15 837.696)	
baths	−10 142.810	
	(22 621.590)	
half_bath	102 491.848	105 609.130
	(26 985.796)	(26 118.559)
sqft	407.408	400.046
	(19.601)	(12.175)
days_on_market	−670.417	−663.538
	(163.646)	(162.787)
hoa_month	131.176	131.373
	(21.511)	(21.487)
property_age	−2415.871	−2367.412
	(358.420)	(344.105)
Num.Obs.	952	952
R2	0.714	0.713
R2 Adj.	0.711	0.711
RMSE	333 693.77	333 736.71

B.2 Alternate Model Summaries

Table 4 presents the model summaries for both the full and reduced models. The full model includes all the variables in the analysis data, while the reduced model excludes the beds and baths variables. The coefficients of the two models are very similar, but the standard errors of these coefficients are slightly lower in the reduced model, suggesting lower uncertainty in the valuations. Regarding the goodness-of-fit metrics, the R^2 scores are also very similar, though the RMSE has increased. However, the reduced number of variables in the latter model results in adjusted R^2 scores that are essentially the same.

C Surveys, Sampling, and Observational Data

This study uses observational data to estimate the intrinsic value of properties in Seattle based on specific parameters. However, observational data has limitations that can reduce valuation accuracy. For instance, deviations from the line of equality in Figure 5 highlight discrepancies arising from factors not captured in the model. Some of these factors are difficult or impossible to measure through observational data alone, whether due to privacy concerns, challenges in data collection, or insufficient information. Surveys provide a valuable way to address these gaps by collecting additional data, which can help reduce the uncertainty in the model outlined in Section 3. For example, surveys could capture qualitative and subjective information, such as owners' expectations of future market trends or personal motivations for selling, that cannot be inferred from observational data.

Relying solely on appraiser data presents another challenge—it often lacks the depth and variety that independent surveys can offer. Appraiser datasets generally focus on tangible features of the property (e.g., square footage, number of rooms) but overlook contextual or personal factors. Independent surveys can fill this gap by gathering details such as a seller's financial situation, their previous purchase price, or even their current perception of market conditions. Additionally, surveys could include questions about less commonly recorded property features like energy efficiency upgrades, aesthetic renovations, or even historical significance, which could significantly influence valuation. By incorporating such data into the model, a better understanding of the factors shaping a property's market value can be gained. This not only allows for more precise valuations but also helps assess the behavioral factors affecting seller decisions, which appraiser data alone cannot address.

Sampling is a vital component of the survey process, as it determines the reliability and representativeness of the responses. A well-designed sampling strategy minimizes bias and ensures diverse data collection. Simple random sampling is ideal in many cases because it reduces systematic bias and ensures all properties have an equal chance of inclusion. This method is particularly feasible given the size of the Seattle property dataset. However, alternative approaches like stratified sampling might be necessary if the dataset has significant variability across neighborhoods or property types. For instance, stratified sampling could ensure equal representation of high-value and low-value neighborhoods, or newer and older properties, which might exhibit different characteristics.

Additionally, expanding the sampling pool to include properties not currently listed on the market can improve the analysis. Data from these properties could offer insights into external factors like neighborhood desirability, local amenities, or proximity to planned infrastructure projects, as well as internal factors like deferred maintenance or owner sentiment about selling. A multi-stage sampling process could be employed, where neighborhoods are first selected based on certain characteristics (e.g., median income levels or recent development activity), followed by a random selection of properties within those neighborhoods. Responses from this expanded sample would help tune the model to reflect a broader range of scenarios, ultimately improving its robustness and applicability.

The survey will be administered in person, where surveyors will knock on each household in the sample's door to ensure detailed and accurate data collection, allowing interviewers to clarify questions. Participants will include property owners, both those currently selling and those who are not, to gather diverse perspectives. The survey will feature questions such as:

- "What year did you purchase the property, and at what price?"
- "Have there been any significant renovations or upgrades?"
- "Is the property fully furnished?"
- "What motivated you to sell (if applicable)?"
- "How would you rate your neighborhood on factors like safety, accessibility, and amenities?"
- "What price range do you believe is fair for this property?"

These questions aim to uncover financial, sentimental, and qualitative factors influencing property valuation. This survey shares similarities with the American Housing Survey, which is another household survey that interviews all types of residents such as property managers, regular occupants, etc. (Bureau 2021)

Challenges to measurement include potential biases, such as response bias, where respondents may provide answers they perceive as favorable rather than accurate (e.g., overstating the quality of renovations). Additionally, respondents may have incomplete or inaccurate records of past transactions or upgrades. Interviewer bias is another concern, as the phrasing or tone of questions might unintentionally influence responses. Addressing these challenges involves careful question design, interviewer training, and cross-referencing self-reported data with available public records where possible.

References

Abelson, Peter W. 1979. "Property Prices and the Value of Amenities." *Journal of Environmental Economics and Management* 6 (1): 11–28. https://doi.org/https://doi.org/10.1016/0095-0696(79)90018-4.

Bankrate. 2024. "What Is the MLS? Multiple Listing Service, Explained." https://www.bankrate.com/real-estate/mls-multiple-listing-service/.

Bureau, U. S. Census. 2021. "American Housing Survey." https://www.census.gov/programs-surveys/ahs.html.

Coulson, N. Edward, and Daniel P. McMillen. 2008. "Estimating Time, Age and Vintage Effects in Housing Prices." *Journal of Housing Economics* 17: 138–51. https://doi.org/10.1016/j.jhe.2008.03.002.

Din, Allan, Martin Hoesli, and André Bender. 2001. "Environmental Variables and Real Estate Prices." *Urban Studies* 38 (11): 1989–2000. https://doi.org/10.1080/00420980120080899.

Estate, Weisbarth Real. 2024. "Seattle Real Estate Market Trends in 2024: What Buyers Should Know." https://www.weisbarth.com/post/seattle-real-estate-market-trends-in-2024-what-buyers-should-know#:~:text=For%20years%2C%20Seattle%20has%20been,growth%20seen%20in%20previous%20years.

Frick, Hannah, Fanny Chow, Max Kuhn, Michael Mahoney, Julia Silge, and Hadley Wickham. 2024. *Rsample: General Resampling Infrastructure*. https://CRAN.R-project.org/package=rsample.

Gao, Shi, Pettit, and Han. 2022. "Property Valuation Using Machine Learning Algorithms on Statistical Areas in Greater Sydney, Australia." *Land Use Policy* 123: 106409. https://doi.org/10.1016/j.landusepol.2022.106409.

Jones, Jonathan. 2024. "Cities with Highest Home Price-to-Income Ratios." https://constructioncoverage.com/research/cities-with-highest-home-price-to-income-ratios.

Kim, Ikhan. 2021. "Spatial Distribution of Neighborhood-Level Housing Prices and Its Association with All-Cause Mortality in Seoul, Korea (2013-2018): A Spatial Panel Data Analysis." *SSM - Population Health* 16: 100963. https://doi.org/10.1016/j.ssmph.2021.100963.

List, Apartment. 2024. "Townhouse Vs Apartment." https://www.apartmentlist.com/renter-life/townhouse-vs-apartment.

National Association of Realtors. 2024. "What to Know about the Appraisal Process." https://www.nar.realtor/magazine/tools/client-education/handouts-for-buyers/what-to-know-about-the-appraisal-process.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Realtors, National Association of. 2024. "Remodeling Impact Report: Outdoor Features." https://www.nar.realtor/research-and-statistics/research-reports/remodeling-impact-report-outdoor-features.

Redfin. 2024. "Seattle, WA homes for sale & real estate." https://www.redfin.com/city/16163/WA/Seattle.

Sarip, Hafez, and Daud. 2016. "Application of Fuzzy Regression Model for Real Estate Price Prediction." *Malaysian Journal of Computer Science* 29 (1): 15–27. https://doi.org/10.22452/mjcs.vol29no1.2.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Yan, Yachen. 2024. *MLmetrics: Machine Learning Evaluation Metrics*. https://CRAN.R-project.org/package=MLmetrics.

Yu, Shuyi. 2020. "Algorithmic Outputs as Information Source: The Effects of Zestimates on Home Prices and Racial Bias in the Housing Market."

Zillow. 2024. "What Is a Zestimate?" https://www.zillow.com/z/zestimate/.