Predicting House Prices in King County, USA:
A Data-Driven Analysis

Gulnar Armour
Springboard
December 26, 2023

1. Introduction

This project utilizes machine learning techniques to predict house prices using the "House Sales in King County, USA" dataset obtained from Kaggle. The project focused on building and evaluating machine learning models to accurately forecast house prices based on a comprehensive set of features provided in the dataset. By using advanced predictive modeling, the project sought to deliver valuable insights for real estate stakeholders and prospective homebuyers.

The primary objective of the project was to develop robust machine learning models capable of accurately predicting house prices. I explored various regression techniques and model evaluation strategies to achieve optimal predictive performance. I conducted in-depth data preprocessing, feature engineering, and hyperparameter tuning to ensure the models were well-optimized. Additionally, thorough cross-validation and performance metrics were used to assess the models' accuracy and generalization capabilities. The ultimate goal was to provide stakeholders with reliable and actionable predictions to support informed decision-making in the real estate market.

Dataset Description

The "House Sales in King County, USA" dataset contains 21,613 records and 20 features. The dataset includes the following attributes:

Property Characteristics:

- Bedrooms
- Bathrooms
- Square footage of living space (sqft_living)
- Square footage of the lot (sqft_lot)
- Number of floors
- Waterfront status
- View quality
- Property condition
- Property grade
- Square footage above ground level (sqft_above)

- Square footage of the basement (sqft_basement)
- Year built
- Year renovated

Location Data:

- Zip code
- Latitude
- Longitude

Comparative Features:

- Square footage of living space for the 15 nearest neighboring properties (sqft_living15)
- Square footage of lot area for the 15 nearest neighboring properties (sqft_lot15)

Transaction Information:

- Unique identifier (id)
- Sale date
- Sale price

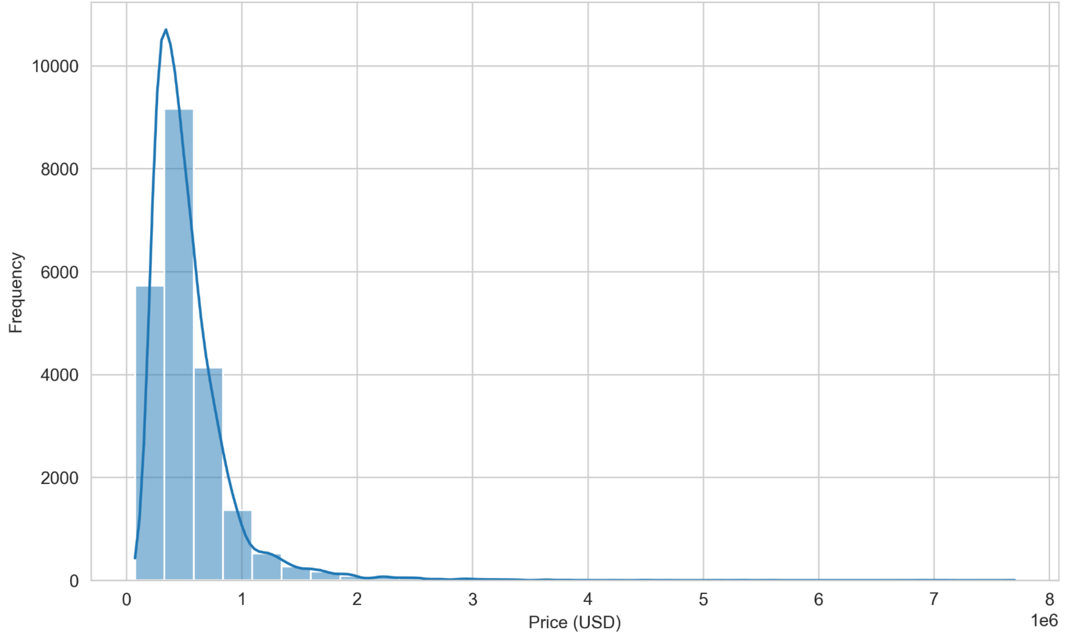
The dataset contains a mix of numerical and categorical features, providing a comprehensive set of attributes for analyzing and predicting house prices in King County, USA.

2. Exploratory Data Analysis

The dataset contains various features related to real estate properties, including 'price', 'bedrooms', 'bathrooms', 'sqft_living', 'sqft_lot', and geographical information such as 'zip code', 'lat', and 'long'. Additionally, I also included some derived features, such as 'log_price', 'log_sqft_living', and 'house_age', as part of my data exploration and preprocessing process. Through this analysis, I gained insights into house pricing dynamics.

Univariate Analysis of House Prices

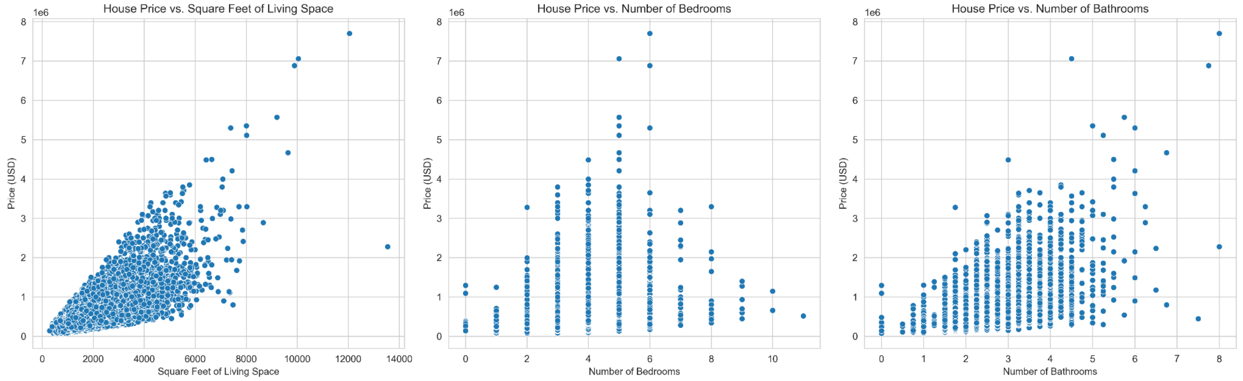
Figure 2.1: Distribution of House Prices



This histogram in Figure 2.1 shows the distribution of house prices in the dataset. The distribution appears right-skewed, indicating that most houses are in the lower price range, with fewer houses in the higher price range. This is a typical distribution for house prices and helps in understanding the range and variance of prices we are dealing with.

Bivariate Analysis: House Price vs. Key Features

Figure 2.2: House Price vs. Property Features


House Price vs. Square Feet of Living Space

House Price vs. Number of Bedrooms

House Price vs. Number of Bathrooms

1. House Price vs. Square Feet of Living Space

This scatter plot shows a positive correlation between the square footage of living space and the house price. Generally, as the living space increases, the price of the house tends to increase. This trend is expected as larger homes typically command higher prices.

2. House Price vs. Number of Bedrooms

The relationship between the number of bedrooms and house price shows a positive trend, but with more variability compared to living space. While generally, more bedrooms correspond to higher prices, the relationship is less linear, indicating other factors also significantly influence price.

3. House Price vs. Number of Bathrooms

Similar to bedrooms, more bathrooms tend to be associated with higher house prices. However, this relationship also displays significant variability, suggesting that while the number of bathrooms is an important factor, it's not the sole determinant of price.

Correlation Matrix of House Features

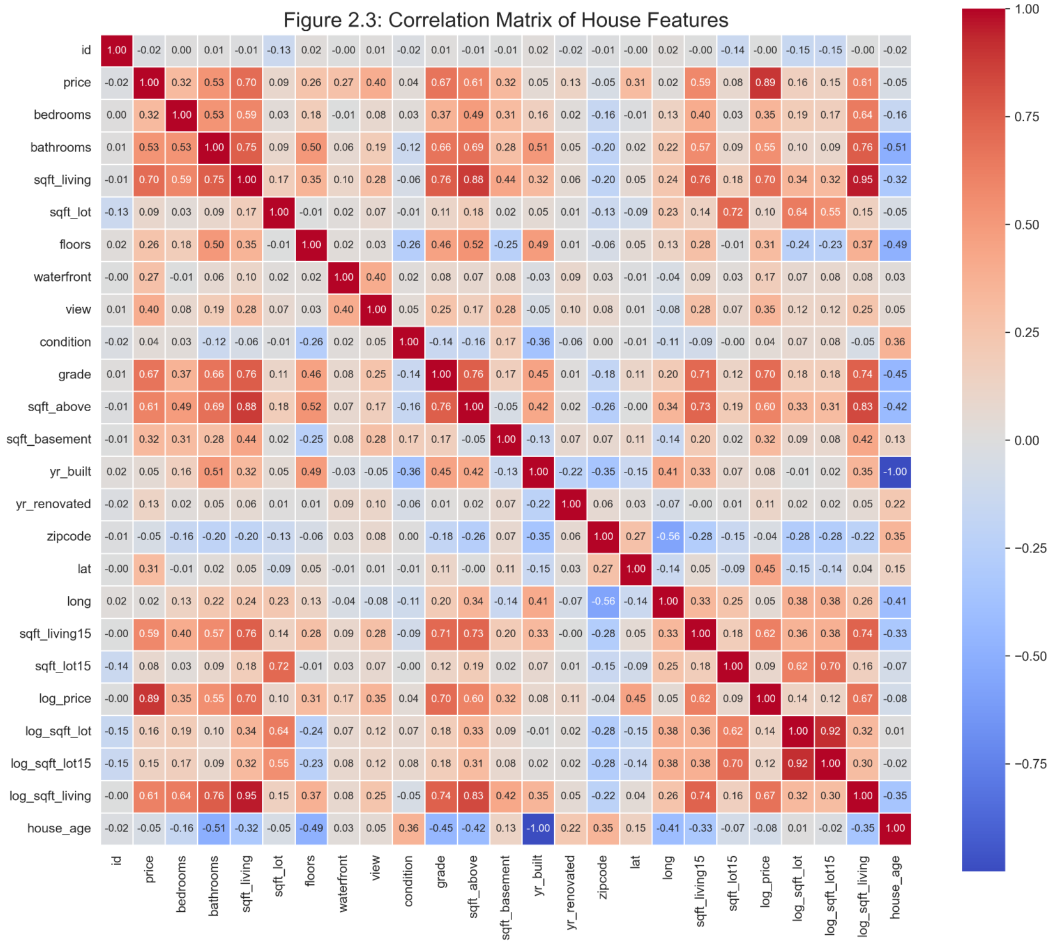


Figure 2.3: Correlation Matrix of House Features

Figure 2.3 shows the correlation coefficients among the dataset's features.Darker red indicates a stronger positive correlation, and darker blue indicates a stronger negative correlation.

Multicollinearity: We observe some features showing high positive correlation with each other, which might indicate multicollinearity. For instance, 'sqft_living' might be highly correlated with 'bedrooms' and 'bathrooms', suggesting these features provide overlapping information.Correlation with Target (Price): Some features show a strong positive correlation with the house price, like 'sqft_living', indicating they are significant predictors of house price.

Statistical Analysis

To delve into a more detailed statistical analysis, I conducted hypothesis tests for the relationships observed in the bivariate analysis. This helped me determine if the observed relationships between house prices and features like 'sqft_living', 'bedrooms', and 'bathrooms' were statistically significant.

Table 2.1: The Pearson correlation

Feature	Correlation Coefficient	P-Value	Result
Square Feet of Living Space	0.702	<0.001	Statistically Significant
Number of Bedrooms	0.315	<0.001	Statistically Significant
Number of Bathrooms	0.525	<0.001	Statistically Significant

The Pearson correlation tests, as shown in Table 2.1, revealed significant relationships between house prices and several key factors. The analysis found a strong and significant correlation (0.702) between the square feet of living space and house price, indicating that larger living spaces are associated with higher property values. Additionally, the number of bedrooms and bathrooms also showed statistically significant correlations with house prices, although the relationships were relatively weaker compared to square footage. The positive correlation (0.315) between the number of bedrooms and house price suggests that bedrooms impact property prices, but other factors also play a role. Similarly, the moderately strong correlation (0.525) between the number of bathrooms and house price indicates that properties with more bathrooms tend to have higher values. These findings emphasize the importance of these property features in influencing house prices, providing valuable insights for real estate professionals and stakeholders in the housing market.

3. Feature Engineering

In this phase of the project, I focused on enhancing the dataset for my model on "Predicting House Prices in King County, USA." This involved not only crafting new features but also transforming existing ones, including the target variable, to better capture the underlying patterns and relationships. The key features engineered and transformations applied are as follows:

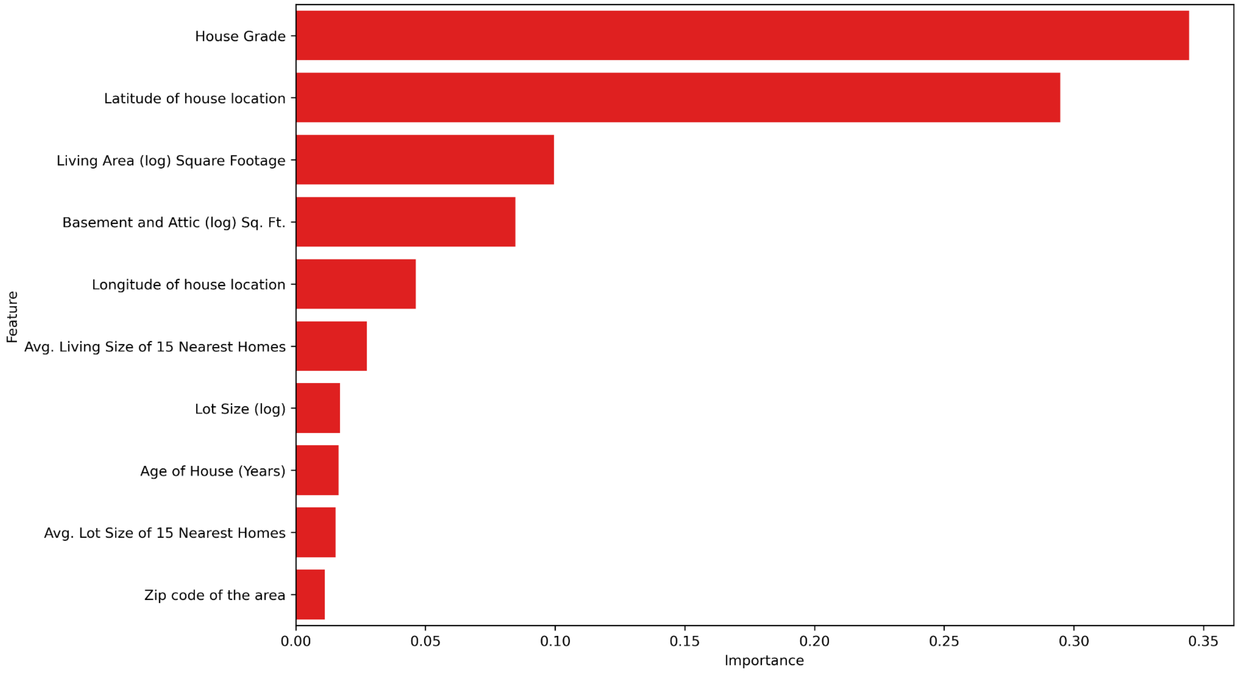
- **Target Variable Transformation (Price):**
    - Transformation: Applied a logarithmic transformation to the target variable, 'price'.
    - Purpose: To address its right-skewed distribution. This transformation helps in stabilizing the variance and improving the model's ability to interpret the data more accurately.
- **sqft_extra:**
    - Creation: This feature was derived by summing 'sqft_above' and 'sqft_basement'.
    - Purpose: To amalgamate above-ground living space and basement areas, providing a more comprehensive view of a property's total living space.
- **house_age:**
    - Creation: Calculated by subtracting the year the house was built from the year it was sold.
    - Purpose: To quantify the age of the property at the time of sale, crucial for understanding how the age impacts its value.
- **Logarithmic Transformations on Square Footage Variables:**
    - Application: Implemented on 'sqft_living', 'sqft_lot', and the newly engineered 'sqft_extra'.
    - Purpose: To normalize these variables, reducing skewness and enhancing the model's ability to capture non-linear relationships.

By meticulously engineering these features and transforming the target variable, I aimed to build a robust foundation for the predictive models. This approach was designed to ensure that the models not only capture the nuances of real estate valuation in King County but also deliver reliable and interpretable predictions.

4. Feature importances

Figure 4.1 illustrates the top 10 features that significantly influence house prices in King County, USA, as determined by the Random Forest model. The most influential feature, 'House Grade', reflects the overall quality and construction standards of the houses, underscoring the importance that buyers place on the build quality. The 'Latitude of house location' emerges as the second most critical factor, indicating that the north-south positioning of a property plays a pivotal role in its market value. This is closely followed by the 'Living Area (log) Square Footage', emphasizing that larger living spaces, when adjusted for skewness through logarithmic transformation, substantially affect pricing.

Figure 4.1: Top 10 Feature Importance



The 'Basement and Attic (log) Sq. Ft.' feature, which also underwent a logarithmic transformation, ranks notably high in importance, suggesting that additional living spaces like basements and attics are key considerations for homebuyers. The 'Longitude of house location' also figures prominently, although it has a slightly lesser impact than latitude, pointing to the east-west positioning's role in the value assessment.

This feature importance ranking, visualized in Figure 4.1, provides valuable insights into the housing market's dynamics, enabling stakeholders to understand which property characteristics are most valued in the region's real estate landscape.

5. Predictive Modeling

Model Selection

The selection of an appropriate model for predicting house prices in King County was a critical step in my analysis. The goal was to find a model that not only provides accurate predictions but also handles the characteristics of the given data effectively. To this end, I evaluated several models, considering both non-logarithmic and logarithmic features to determine the best performer based on the Root Mean Square Error (RMSE) metric. The summary of my findings drawn from the RMSE values, as detailed in Table 5.1.

Table 5.1: The Root Mean Square Error (RMSE) Values

Model	RMSE with Non-Logged Features	RMSE with Logged Features
Linear Regression	212,242.35	198,835.51
Random Forest	147,732.82	**139,035.95**
Decision Tree	206,839.73	183,689.14
Gradient Boosting	149,466.78	145,788.72

The comparative analysis of RMSE values clearly indicated that the **Random Forest model** was superior in predicting house prices accurately. Its ability to handle both non-logarithmic and logarithmic features effectively, coupled with the lowest RMSE scores, solidified its position as my model of choice. The improvement in RMSE with logarithmic features across all models confirmed my hypothesis that normalizing data distributions leads to better model performance.

Cross-Validation for Random Forest

To further validate the robustness of my Random Forest model, I conducted a 5-fold cross-validation. This method entailed dividing the dataset into five distinct parts, training the model on four of these parts, and then evaluating it on the remaining part, for each of the five folds.

Table 5.2: 5-fold Cross-Validation

Fold Number	RMSE (USD)
1	137,847.476
2	127,117.2315
3	118,312.9399
4	127,756.5731
5	129,509.7377
Mean	128,108.7916
Std Dev	6,226.8866

Variability Across Folds:

The RMSE values across the 5 folds range from approximately 118,313 to 137,848 USD. This variation indicates how the model's performance fluctuates with different subsets of data.

Average Performance:

The mean RMSE is 128,109 USD, suggesting that, on average, the model's predicted house prices are within this range from the actual prices. Given the context of house pricing, this average error can be considered in relation to the overall price range and distribution of house prices in the dataset.

Consistency of the Model:

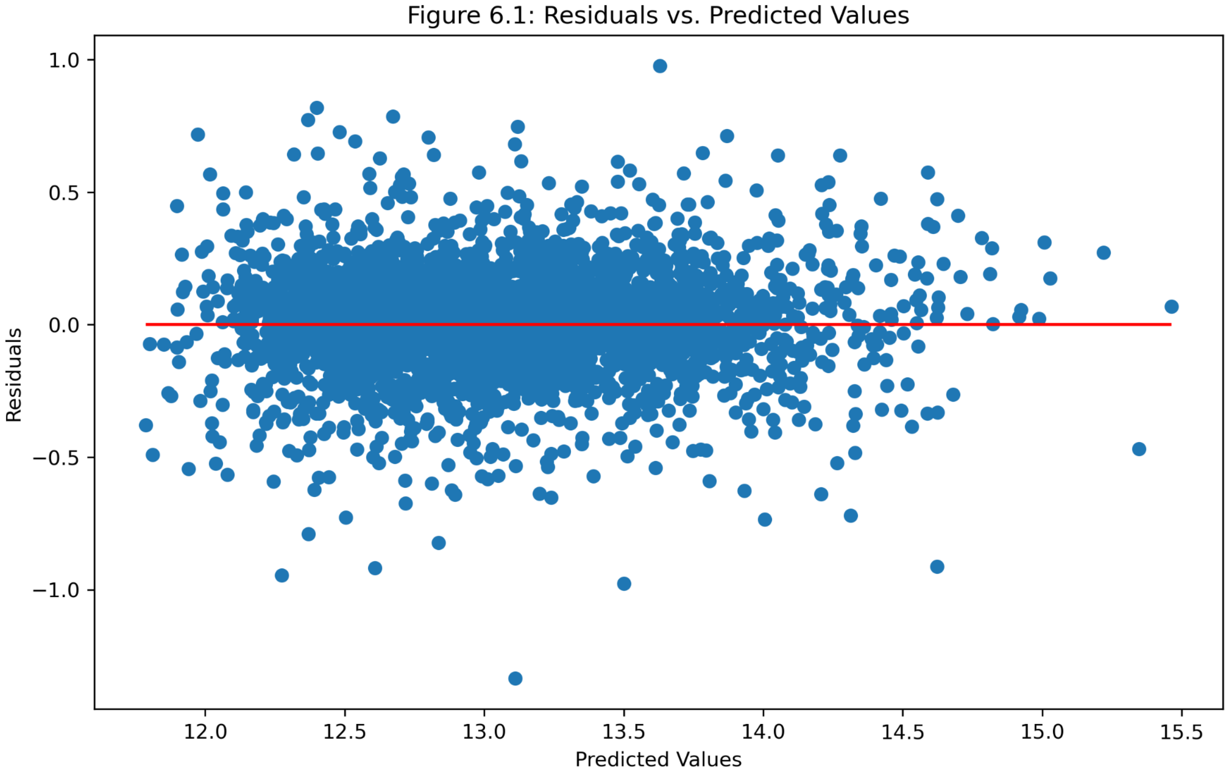
The standard deviation of the RMSE scores is approximately 6,227 USD. This relatively small value, in comparison to the mean RMSE, indicates that the model's performance is consistent across different data subsets.

While the model shows reasonable consistency, the decision to use this model should be based on how the mean RMSE aligns with the specific business objectives and the acceptable error margin in predicting house prices. For instance, in high-stakes scenarios like real estate investment or lending, even a small percentage of error could translate to a significant monetary impact, warranting a more stringent evaluation of the model's accuracy.

6. Model Evaluation

Residual Analysis

Reflecting on the residual plot from the predictive model, depicted in Figure 6.1, a pattern is observed that's largely in line with what I would hope to see — the residuals, representing the differences between the observed and predicted house prices, cluster around the zero line. This suggests that the model, on average, is making predictions that are neither systematically overestimated nor underestimated.



The scatter of points is relatively even across the range of predicted values, though there's a slight increase in spread as the predicted prices rise. This indicates that for more expensive houses, the model's predictions become less precise. A few points stray far from the zero line, especially at higher predicted values, hinting at outliers or possible leverage points that could be skewing the predictions.

These outliers are intriguing to me; they represent cases where the model's predictions deviate significantly from the actual values. I'm particularly keen to investigate these further, as they may uncover data entry errors, rare property features not captured by the dataset, or even new trends emerging in the housing market.

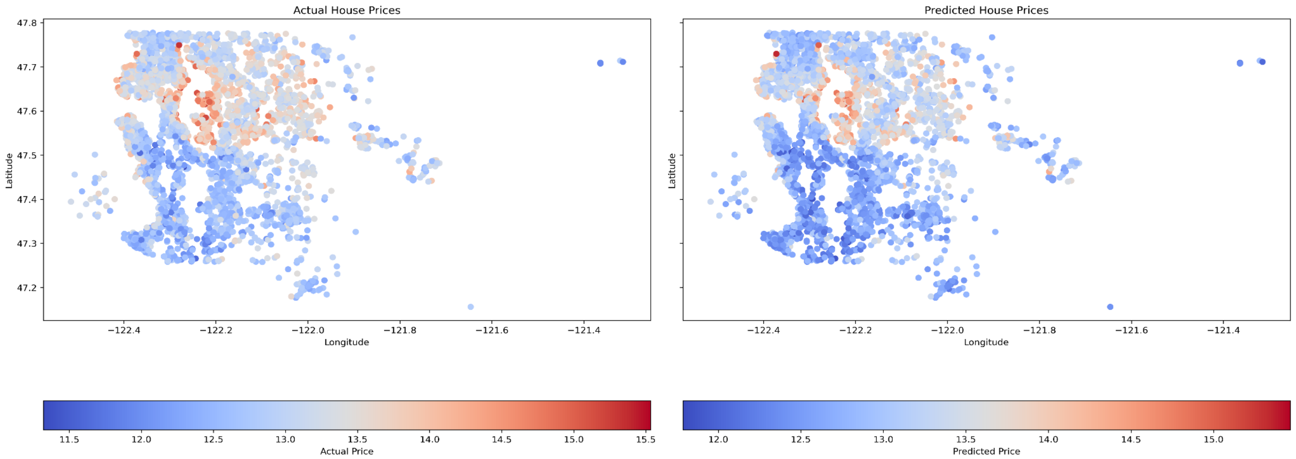
While the plot doesn't show any obvious systematic patterns — no clear curves or trends — I note that the assumption of homoscedasticity seems reasonable here. The variance of the residuals doesn't appear to increase or decrease with the predicted prices, suggesting consistent prediction errors across the price spectrum.

This residual plot serves as a crucial diagnostic tool, reaffirming the robustness of the model while also highlighting areas that may benefit from further refinement. It prompts me to consider deeper analyses, particularly looking into the high-value property predictions, to understand and improve upon the model's performance.

Geospatial Validation

After examining the side-by-side geospatial validation plots I created in Figure 6.2, I am impressed by how accurately the predictive model reflects the real spatial distribution of house prices. The left plot, depicting actual prices, reveals a vibrant mosaic of values across the region, with pockets of intense reds indicating high-value properties interspersed among cooler blues. The right plot, illustrating predicted prices, echoes this pattern with a remarkable degree of visual congruence, affirming the model's effectiveness at capturing the geographic nuances that shape property valuation.

Figure 6.2: Geospatial Distribution of Actual vs Predicted House Prices



While the model adeptly identifies areas of both affluence and modesty, a closer inspection suggests a slight smoothing effect in the predicted prices. This is most notable in the scatter of the highest-priced homes; my model appears to temper the extremes, suggesting a conservative estimation on the upper end of the price spectrum. This could be a reflection of the model's inherent design, which aims to generalize and may thus be less sensitive to the outliers that can characterize luxury markets.

The comparison offers a compelling narrative of spatial consistency, yet it also highlights opportunities for refinement. The predictive model, while robust, could benefit from incorporating additional features—perhaps those capturing unique architectural elements, historical significance, or proximity to coveted amenities—that may influence the upper echelons of the market.

In conclusion, these geospatial plots serve as a testament to the model's validity, showcasing a strong alignment with the real-world data. They also chart a course for future explorations, where deeper granularity and an enhanced feature set could further sharpen the model's precision, particularly in predicting those rarefied properties that defy the norm.

7. Conclusion

As I conclude this project, I reflect on the process that has enabled me to gain a deeper understanding of the factors influencing house prices. My analysis, grounded in robust statistical methods and enriched with geospatial insights, has provided me with a clearer picture of the real estate market dynamics.

Through meticulous model selection and evaluation, I discovered that the Random Forest model, with its ensemble approach, yielded the most accurate predictions among the various algorithms I explored. The cross-validation process reinforced my confidence in its reliability, and the residual analysis offered further validation of its predictive power.

The feature importance analysis was particularly revealing, highlighting house grade and latitude as pivotal factors in determining house prices. This insight has significant implications for stakeholders in the real estate industry, suggesting a strong regional component to property valuation and the critical role of property location.

The geospatial validation brought an additional layer of understanding, allowing me to visualize how well the model's predictions corresponded with actual market prices across different locations. This spatial component of my analysis not only confirmed the model's effectiveness but also provided a tangible connection between data-driven predictions and the real-world landscape where these properties exist.

In summary, this project has affirmed the value of advanced machine learning techniques in interpreting complex datasets and has underscored the importance of quality, location and size in the real estate market. As I move forward, I am eager to apply these insights and methodologies to new challenges, continually refining my models to capture the nuances of property valuation. Future work may involve exploring additional features, implementing more sophisticated models, and expanding the dataset to include temporal factors for a dynamic market analysis.

This endeavor has not only honed my skills in data science but also enriched my understanding of the intricate interplay between various features and house prices. I am optimistic that the methodologies and findings of this project will be a valuable asset for decision-makers in the real estate sector, guiding strategic investments and policy development.

8. Appendix

Data Dictionary

Feature	Description	Non-null Count	Data Type
id	Unique identifier for each house	21613	Integer
date	Date the house was sold	21613	Object
price	Price of the house	21613	Float
bedrooms	Number of bedrooms	21613	Integer
bathrooms	Number of bathrooms	21613	Float
sqft_living	Square footage of the living area	21613	Integer
sqft_lot	Square footage of the lot	21613	Integer
floors	Number of floors in the house	21613	Float
waterfront	Whether the house has a waterfront view	21613	Integer
view	Number of views from the house	21613	Integer
condition	Overall condition of the house	21613	Integer
grade	Overall grade given to the housing unit	21613	Integer
sqft_above	Square footage of house apart from basement	21613	Integer
sqft_basement	Square footage of the basement	21613	Integer
yr_built	Year the house was built	21613	Integer
yr_renovated	Year the house was last renovated	21613	Integer
zipcode	Zip code of the area	21613	Integer
lat	Latitude of the house location	21613	Float

long	Longitude of the house location	21613	Float
sqft_living15	Avg. Living Size of 15 Nearest Homes (sqft)	21613	Integer
sqft_lot15	Avg. Lot Size of 15 Nearest Homes (sqft)	21613	Integer
log_sqft_extra	Total square footage of basement and attic	21613	Integer

Data Source: House Sales in King County, USA