

UCLA

UCLA Electronic Theses and Dissertations

Title

Predictive Modeling of U.S. Apartment Rent Prices

Permalink

<https://escholarship.org/uc/item/51k8d8gw>

ISBN

9798190937367

Author

Wakamoto, Linda

Publication Date

2026-09-08

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Predictive Modeling of U.S. Apartment Rent Prices

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

Linda Wakamoto

ABSTRACT OF THE THESIS

Predictive Modeling of U.S. Apartment Rent Prices

by

Linda Wakamoto

Master of Applied Statistics and Data Science

University of California, Los Angeles, 2026

Professor Schoenberg, Frederic Roland, Chair

Driven by escalating living costs in the United States, this thesis investigates the key determinants of monthly apartment rental prices and evaluates predictive machine learning models using a national apartment listing dataset from the UC Irvine Machine Learning Repository. Exploratory data analysis and hypothesis testing revealed that spatial geography is overwhelmingly the most powerful driver of rental variation, with coastal states like Hawaii, New York, and California having significantly higher prices. Structural layout attributes, specifically the number of bathrooms and bedrooms, also heavily influenced valuations, outperforming listing-level administrative variables like upfront fees and pet policies in predictive power. Preprocessing involved scaling, one-hot encoding, and utilizing the Interquartile range method to eliminate extreme price anomalies. Of all trials, the best predictive models targeted price per square foot, and a Random Forest regressor yielded the highest accuracy on the national scale, while a XGBoost model achieved superior performance for the California listings model. The best prediction models were also used to identify listing outliers that are exceptionally overpriced or that appear to be exceptional values. The study's limitations include geographic sampling bias and feature dilution in tree-based models caused by multi-collinearity from engineered spatial interaction terms. Future research should utilize spatial clustering algorithms (such as K-Means or DBSCAN) to establish explicit regional boundaries, apply algorithmic balancing frameworks like SMOTER to fix regional skewness, and implement text-mining pipelines to extract unstructured data from listing descriptions to capture crucial interior property conditions independent of geographic location.

The thesis of Linda Wakamoto is approved.

Wu, Yingnian

Leong, Oscar

Schoenberg, Frederic Roland, Committee Chair

University of California, Los Angeles

2026

*To my family and friends. . .
and everyone who supported me
during my time at UCLA*

TABLE OF CONTENTS

1	Introduction	1
2	Previous Work	3
3	Data	6
4	Data Preprocessing	7
5	Exploratory Data Analysis	13
6	Modeling	27
7	Conclusion	42
	References	45

LIST OF FIGURES

5.1	Number of Apartment Listings by State	13
5.2	Correlation Heatmap of Property Features for National Apartment Listings .	14
5.3	Distribution of Monthly Rental Prices Across National Listings	15
5.4	Distribution of the Number of Bathrooms per Apartment	16
5.5	Distribution of the Number of Bedrooms per Apartment	16
5.6	Comparison of Monthly Rental Prices for Apartments With and Without Upfront Fees	17
5.7	Comparison of Monthly Rental Prices Based on Dog Allowance Policies . . .	17
5.8	Average Monthly Rental Price Ranked by U.S. State	19
5.9	Geographic Distribution of Mean Apartment Rental Prices Across the U.S. by Price Tier	19
5.10	Correlation Heatmap of Property Features for California Listings	20
5.11	Distribution of Monthly Rental Prices for California Apartments	21
5.12	Distribution of the Number of Bathrooms in California Apartments	22
5.13	Distribution of the Number of Bedrooms in California Apartments	22
5.14	Comparison of Rental Price Distributions for California Apartments With and Without Upfront Fees	23
5.15	Comparison of Rental Price Distributions for California Apartments Based on Dog Allowance	23
5.16	Comparison of Rental Price Distributions for California Apartments and Non- California Apartments	24
5.17	Geographic Distribution of Mean California Apartment Rental Prices by Price Tier	26
6.1	Model Residuals: Best Model for U.S. Apartments	39

6.2	Model Residuals: Best Model for CA Apartments	40
-----	---	----

LIST OF TABLES

6.1	Model Performance Excluding Price Outliers for U.S. Listings	27
6.2	Model Performance Including Price Outliers for U.S. Listings	28
6.3	Model Performance Including Price per Square Foot Outliers for U.S. Listings	30
6.4	Model Performance Excluding Price per Square Foot Outliers for U.S. Listings	31
6.5	Model Performance Including Price per Square Foot Outliers for CA Listings	33
6.6	Model Performance Excluding Price per Square Foot Outliers for CA Listings	34
6.7	Model Performance Including Price per Square Foot Outliers for CA Listings with Distance to Nearest Beach	35
6.8	Model Performance Excluding Price per Square Foot Outliers for CA Listings with Distance to Nearest Beach	36

ACKNOWLEDGMENTS

(Acknowledgments omitted for brevity.)

CHAPTER 1

Introduction

In recent years, macroeconomic pressures in the United States have driven significant inflation and an increase in the cost of living. The housing sector has also been heavily impacted by these trends, as it is increasingly difficult to purchase a home, particularly in California, as studies show that California homes are about twice as expensive as the typical US home [2]. There is a relatively high demand for apartments in California, as Zillow, an apartment listing website, shares that the market temperature for apartments in California is consistently warm, compared to the national average [14].

As young adults face this difficult housing market while they start their careers, a data-driven predictive model for apartment rent costs becomes an essential tool. This thesis utilizes national apartment listings data from the UC Irvine Machine Learning Repository. Prior to predictive modeling, comprehensive exploratory data analysis and statistical testing were conducted to identify trends and relationships between listing characteristics. This thesis aims to answer the question: how can we best predict an apartment's monthly rent prices from rental posting features for the United States and specifically within California?

To answer this question, comprehensive exploratory data analysis and statistical testing were first conducted to identify trends and relationships between listing characteristics. Next, machine learning models such as Multilinear Regression, Random Forest Regressor, XGBoost, and Gradient Boost were trained on different feature sets to predict rental price. These variations included just using numeric feature variables or a mix of categorical and numeric feature variables, filtering by California listings or using the entire dataset, and including and excluding price variable outliers as the target variable. In addition, price per square foot was used as a target variable, and new engineered variables, such as distance to

the nearest beach, were introduced. After performing this predictive modeling and recording the evaluation metrics, the best fitting models were found. Finally, regression analysis was applied to the best models for California listings and all listings to find high and low outliers that show the most overpriced and best underpriced listings.

CHAPTER 2

Previous Work

Macroeconomic and Regional Drivers

When analyzing apartment markets, research often focused on the high-level economic and geographic factors that set baseline prices. For instance, research on Swedish and German apartment prices by Engerstam (2021) has found that macroeconomic fundamentals, including income levels, interest rates, and housing supply, drive long-run apartment price dynamics, with differing responsiveness across markets indicating varying levels of volatility and adjustment [5].

On a more localized level, research on new Prague apartment prices by Mazáček and Panoš (2025) showed that property-specific attributes (such as size, age, and layout) and neighborhood characteristics (location, accessibility, and local amenities) are consistently identified as key determinants of housing values, making them essential inputs for rent prediction models [9]. Similarly, a study done on the effect of priced parking on property prices in Stockholm suburbs by Bondemark and Merkel (2023) showed that amenities such as parking availability positively affect housing prices, with empirical results indicating that accessibility and infrastructure-related variables significantly contribute to property valuation [1].

Frameworks for Feature Analysis

Building on these foundational attributes, researchers have developed structured frameworks to better capture how individual features interact. Manaswini et al. (2024) found that machine learning techniques can effectively model the complex relationships between housing

features and rental prices, enabling more accurate and data-driven rent estimation systems [8].

For rental properties, by contrast, Chiwhane et al. (2023) report that a data-driven framework using machine learning such as Long Short-Term Memory (LSTM) models with social network rental house prices can optimize rental housing decisions by identifying key factors influencing rent and improving predictive accuracy through structured feature analysis [4].

Algorithmic Performance and Model Optimization

Once these features are established, the focus shifts to choosing the right algorithmic approach to handle the data. Research predicting rental prices in India by Zhang (2024) shows that advanced machine learning models such as Gradient Boosting and XGBoost, combined with hyperparameter optimization, significantly enhance rental price prediction accuracy compared to baseline regression methods [13].

On the other hand, a study done on Beijing rental data by Cao et al. (2023) shows that among multiple machine learning models applied to large-scale rental datasets, Random Forest achieves superior performance compared to K-nearest neighbors and Gradient Boosting in predicting rent prices due to the ability to handle nonlinear relationships and large feature sets [3]. Rather than relying on a single algorithm, Wang et al. (2023) proposes that ensemble learning approaches, particularly stacking models, outperform individual machine learning algorithms in rental price prediction by combining the strengths of multiple predictive models [12].

Advanced Deep Learning and Spatial Scalability

Finally, as datasets grow larger and more complex, researchers have turned to deep learning architectures to capture intricate patterns. Han and Chun (2025) found that machine learning approaches such as long short-term memory models have been shown to significantly improve Korean apartment price prediction accuracy by capturing complex nonlinear

relationships and interactions between variables, supporting their application to rental price modeling [6].

Furthermore, as data scales up, traditional geographic models face competition. Research by Seya and Shiroi (2019) found that as sample sizes grow, deep learning models approach or match spatial regression in rent prediction accuracy in Japanese housing markets [10].

CHAPTER 3

Data

The data used for this project is a national apartment listing dataset with variables of the listing details such as location, price, number of bathrooms, and whether pets are allowed. It is the “Apartment for Rent Classified” dataset from the UC Irvine Machine Learning Repository [11], and it has 100,000 rows and 21 variables. For this analysis, eight apartment listing variables were used, as well as the price of the apartment listing (monthly rent). The numeric variables that were selected were **bathrooms** (the number of bathrooms), **bedrooms** (the number of bedrooms), **longitude**, and **latitude**. The categorical variables that were selected were **fee** (whether there is an initial fee), **photo** (there is a photo on the listing), **pets_allowed** (if and what kind of pets are allowed), and **state** (the state the listing is located in).

This dataset was donated to UC Irvine on December 25, 2019 and it consists of apartment rental prices from January 2019 through January 2020 from online CA apartment listings. The listings are from 25 different online listing sources, and are mostly from RentDigs.com. For the sake of the time available for this analysis, the variables that were excluded are website source, ID, category, title, body, amenities, currency, address, price type, price display, and time, but these can be used in future analysis to strengthen predictive models, as listed in the conclusion section.

CHAPTER 4

Data Preprocessing

Feature and Target Selection

The raw dataset was filtered to isolate a specific subset of structural, administrative, and geographic features. The final independent feature matrix (X) was restricted to ten primary attributes: `bathrooms`, `bedrooms`, `square_feet`, `fee`, `cityname`, `has_photo`, `pets_allowed`, `state`, `latitude`, and `longitude`.

The dependent target variable (Y) was defined in two ways for different modeling trials: monthly rent (`price`), and the price per square foot (`price_per_square_foot`). The `price_per_square_foot` metric normalizes property valuations across varying spatial dimensions, preventing large luxury properties from disproportionately skewing the loss functions of the estimators. For the localized regional analysis, an identical pipeline was engineered for a filtered subset containing only California observations (`df_CA`), omitting the `state` column to eliminate multi-collinearity.

Data Cleaning and Structural Adjustments

Prior to statistical modeling, extreme anomalies within the rental price distribution were systematically isolated and removed to prevent heavy-tailed distortions. Utilizing the Interquartile Range (IQR) method, extreme price variable outliers were eliminated, resulting in the removal of 4,707 outliers. Following this outlier mitigation step, the target vector and independent feature matrix were synchronized to ensure row alignment.

This processing framework structurally altered the dataset scale: the baseline national

dataframe encompassed 99,759 initial rows across 9 columns, which expanded to 70 total columns following nominal encoding. After filtering out the 4,707 price outliers, the finalized national modeling dataset retained a clean sample size of 95,052 rows, which was used uniformly across all predictive models except for Kernel Regression, where a sampled subset of 5,000 rows was extracted to satisfy computational boundaries. For the targeted regional sub-analysis, filtering the national data exclusively for listings inside California yielded a localized sample size of 10,338 rows.

Imputation Framework for Missing Data

To ensure completeness of the data and satisfy the input constraints of scikit-learn estimators, a systematic imputation pipeline was executed using the `SimpleImputer` architecture. Missing values across both features and targets were treated under a deterministic statistical baseline to maximize the sample size and avoid row-deletion bias.

Prior to executing the imputation step, feature completeness was quantified. The dataset exhibits a high degree of data integrity, as the amount of missing data across all primary features rested well under one percent. Specifically, within the national dataset, the discrete `bedrooms` variable exhibited the highest rate of omission at approximately 0.12%, followed closely by the discrete `bathrooms` variable at 0.06%. The geographic coordinate fields (`latitude` and `longitude`) were missing in only 0.03% of listings, while the continuous target variable `price_per_square_foot` was omitted in a negligible 0.001% of rows. Crucially, physical attributes such as `square_feet` and administrative indicators like `fee`, `has_photo`, and `pets_allowed` were entirely complete with 0.00% missingness.

For the targeted California sub-analysis (`df_CA`), a localized missingness profile emerged where the discrete `bedrooms` attribute constituted the only major anomaly, exhibiting a minor omission rate of 0.52%, while the `bathrooms` variable was missing in 0.03% of cases. All other features—including spatial coordinates, property square footage, and pricing targets—remained completely populated within the regional subset.

Feature Matrix Imputation

The independent variables were partitioned by structural type into numeric and categorical vectors to apply appropriate mathematical adjustments.

Feature Matrix Imputation

The independent variables were partitioned by structural type into numeric and categorical vectors to apply appropriate mathematical adjustments:

- **Numerical Features:** For continuous and discrete numeric variables (`bathrooms`, `bedrooms`, `square_feet`, `latitude`, and `longitude`), missing entries were substituted using a global median strategy. The median was selected over the mean to provide robust central tendency mitigation, ensuring that extreme geographical coordinates or large square feet did not pull the imputed baselines into unrepresentative dimensions.
- **Categorical Features:** For nominal attributes (`fee`, `cityname`, `has_photo`, `pets_allowed`, and `state`), missing metrics were resolved using a `most_frequent` frequency baseline. This mode-based replacement assigns the most statistically common category within the dataset to the missing observation, preserving the modal probability distribution of categorical listing variables.

Numerical Features

For continuous and discrete numeric variables (`bathrooms`, `bedrooms`, `square_feet`, `latitude`, and `longitude`), missing entries were substituted using a global median strategy. The median was selected over the mean to provide robust central tendency mitigation, ensuring that extreme geographical coordinates or large square feet did not pull the imputed baselines into unrepresentative dimensions.

Categorical Features

For nominal attributes (`fee`, `cityname`, `has_photo`, `pets_allowed`, and `state`), missing metrics were resolved using a `most_frequent` frequency baseline. This mode-based replacement assigns the most statistically common category within the dataset to the missing observation, preserving the modal probability distribution of categorical listing variables.

Target Variable Imputation

In alignment with the feature matrix adjustments, missing values within the dependent target variable, `price_per_square_feet`, were similarly resolved using a median imputation framework. This was deployed here to safeguard the integrity of the data boundaries and maintain standard row mapping during full data transformations. To guarantee execution safety prior to modeling, a hard programmatic assertion was embedded into the preprocessing pipeline:

$$\sum \text{Null}(Y) = 0$$

This structural gate enforces that zero missing observations remain in the target vector, preventing runtime failures during gradient descent or matrix multiplication.

Feature Transformation and Multi-Dimensional Encoding

Following imputation, the feature space was compiled using a `ColumnTransformer` to execute parallel numeric scaling and categorical binarization.

Categorical Binarization

The categorical variables were processed using a `OneHotEncoder` mapping for machine learning models. This transformation maps nominal string categories into binary indicator variables (such as expanding categorical states into distinct elements like `state_CA`). To prevent data leakage and handle unexpected strings during future testing phases, the encoder

was configured to ignore unknown categories gracefully (`handle_unknown='ignore'`). This sparse expansion fundamentally altered the dimensionality of the independent feature space, expanding the network from 9 base features up to 70 distinct continuous and boolean parameters.

Continuous Feature Standardization

Because distance-dependent and linear estimators are highly vulnerable to objective magnitude differences between input features (such as comparing a decimal latitude to a four-digit square footage), all numeric features were transformed via a global `StandardScaler`. This operation normalizes each column by subtracting its empirical mean (μ) and scaling it to unit variance through division by its standard deviation (σ):

$$z = \frac{x - \mu}{\sigma}$$

This mathematical alignment centers the distribution of physical features around an isotropic mean of zero, ensuring uniform gradient steps and preventing high-magnitude features from artificially dominating model parameter optimization. The resulting matrices were cast into uniform 64-bit floating-point arrays (`np.float64`) to preserve maximum numerical precision throughout all subsequent predictive trials.

Feature Engineering and Coastal Spatial Metrics

To capture the localized economic influence of geographic proximity to marine environments within the California sub-market, specialized spatial attributes were engineered and appended to the California listings dataset `df_CA`. A geographic distance matrix was generated to compute the shortest distance in kilometers from each individual apartment listing’s coordinates to the nearest coordinate pair in a compiled matrix of major California beaches, producing the continuous feature `dist_to_nearest_beach_km`.

From this baseline proximity field, a secondary binary feature—`is_waterfront`—was

established, mapping an indicator flag of 1 to any property positioned within immediate walking distance (defined as a strict threshold of ≤ 1 km) of the coast, and 0 otherwise. Finally, to control for varying localized demand patterns along the extensive North-South axis of the state, a continuous interaction term (`coastal_lat_interaction`) was formulated by multiplying the beach distance by the property's latitude. This scales the spatial distance metric against regional geography, reflecting the hypothesis that coastal proximity yields distinct premium dynamics in Southern California's climate relative to Northern California real estate markets. These engineered variables were integrated across specific modeling trials to evaluate localized coastal premiums.

CHAPTER 5

Exploratory Data Analysis

4 Exploratory Data Analysis - All States

State Variable Counts

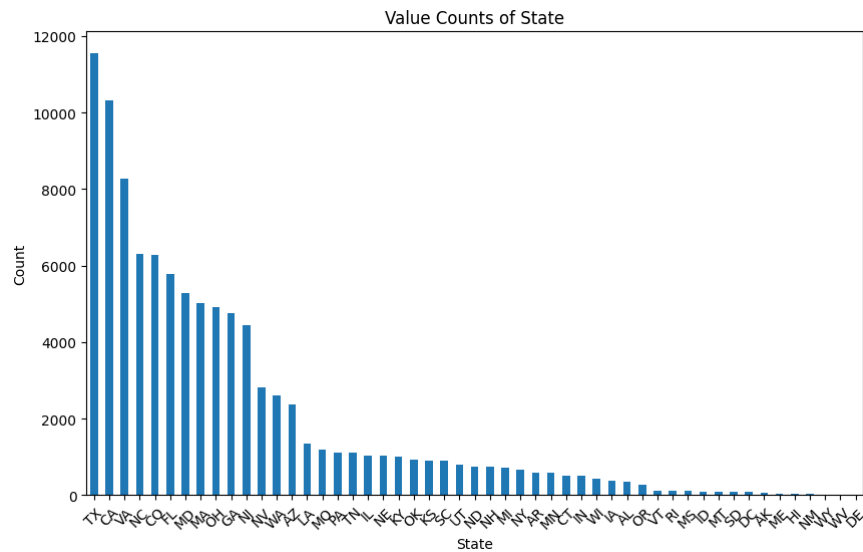


Figure 5.1: Number of Apartment Listings by State

The bar chart in Figure 5.1 (Figure 5.1) shows the distribution of states in the state variable of our dataset. It is apparent that this dataset is heavily skewed, and over-represents states such as Texas, California, and Virginia. This over-representation worked well for my analysis on just California states, but is a clear limitation for my national apartment analysis.

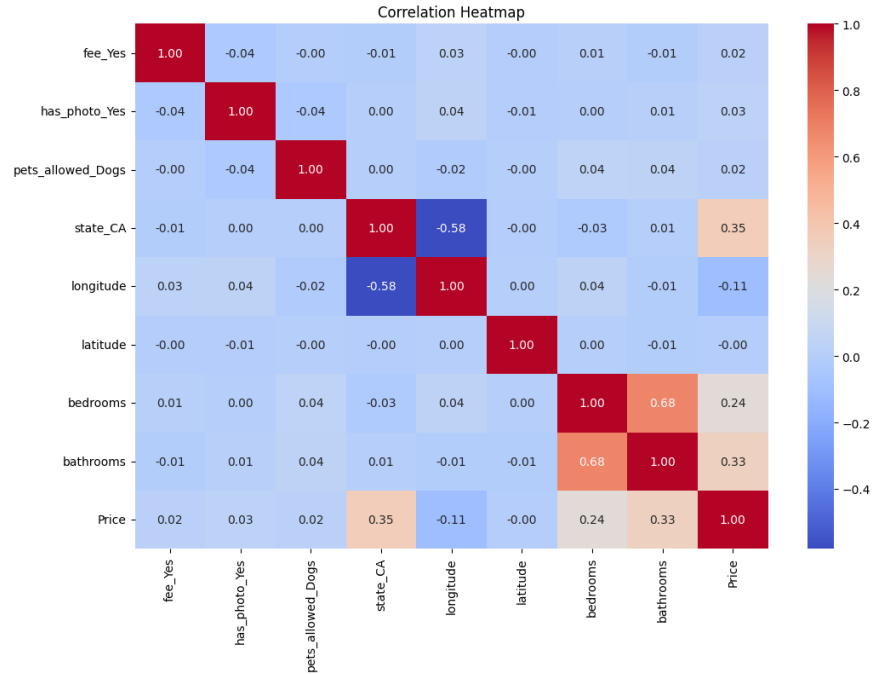


Figure 5.2: Correlation Heatmap of Property Features for National Apartment Listings

Correlation Heatmap

The correlation heatmap in Figure 5.2 (Figure 5.2) shows which variables have a high linear correlation to the price variable, and can inform us of which variables may have high feature importance in predictive modeling.

This heatmap shows that the variables with the highest correlation to price from the ones I chose for this analysis are the listing location in California, the number of bathrooms for the listing, and the number of bedrooms of the listing. The correlation score for the California location variable was 0.35 to price, and the number of bathrooms was 0.33, and the number of bedrooms was 0.24.

The other variables did not have high correlation to price (below 0.11), but they can still have a nonlinear or complex interaction with price. Other high correlations were between longitude and location in CA (-0.58), number of bedrooms and number of bathrooms (0.68).

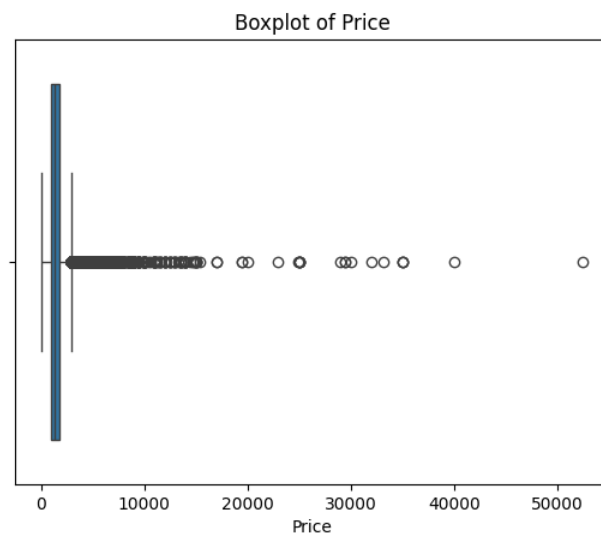


Figure 5.3: Distribution of Monthly Rental Prices Across National Listings

Price

There were many high outliers for price (Figure 5.3), and the median price was \$1,350.00 while the mean price was \$1,527.21 with standard deviation \$903.64. The minimum, \$100/month, was for a listing in Connecticut that had one bathroom, two bedrooms, Cats and Dogs allowed. The listing's latitude was 41.79 and longitude was -72.67. According to Google Earth, this seems to be located in Hartford, Connecticut. The maximum, \$52,500/month, was for a listing in California that had one bathroom and no bedrooms, where Cats and Dogs were allowed, and latitude was 34.89 and longitude was -117.04. According to Google Earth, this is located in Barstow, California.

The mean price per square foot in national listings was \$1.68/sqft (median: \$1.47/sqft) and the minimum was \$0.05/sqft while maximum was \$37.02/sqft. The standard deviation was \$0.85/sqft.

Bathrooms and Price

There were 82 high outliers for the bathroom variable (Figure 5.4). The mean number of bathrooms was 1.73 with a standard deviation of 0.75, and the median was 1. The p-value

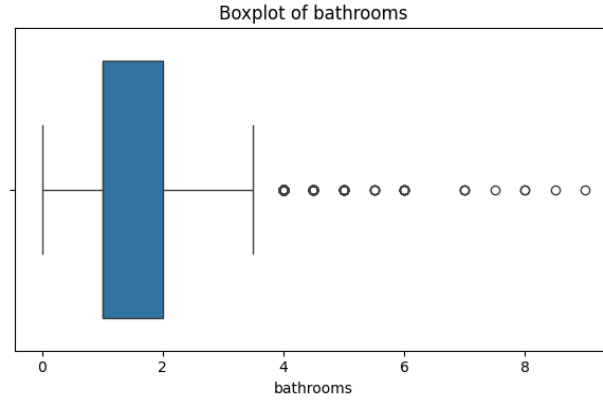


Figure 5.4: Distribution of the Number of Bathrooms per Apartment

was statistically significant for all bathroom comparisons (bathrooms 0 vs 1, 1 vs 2, 2 vs 3, 3 vs 4). The maximum number of bathrooms is nine, and this is for a listing at Lake Hollywood Drive, Los Angeles, CA 90068 [7]. The price for this listing is \$32,000/month.

Bedrooms and Price

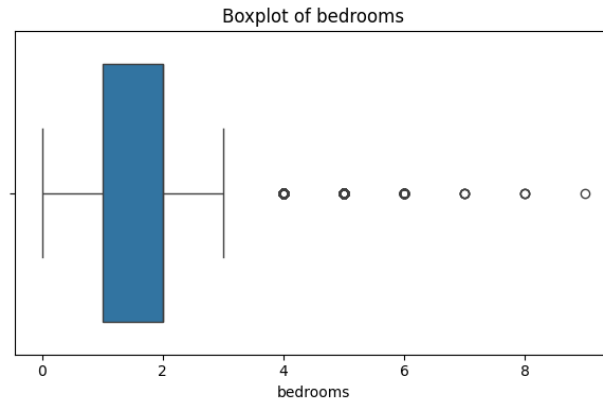


Figure 5.5: Distribution of the Number of Bedrooms per Apartment

There were 341 high outliers for the **bedrooms** variable (Figure 5.5). The mean bedroom was 1.45, with a standard deviation of 0.55, and the median was 2. There is a listing in Kalamazoo, Michigan with nine bedrooms (rent: \$3,159/month), and this is the maximum number of bedrooms from this data. There was a statistically significant difference in mean rent prices between bedroom changes by 1 (for rooms 0 to 4). The maximum number of

bedrooms is nine, and this is for a listing at 310 East Vine Street, Kalamazoo, MI 49001 [7]. The price for this listing is \$3,195/month.

Other Variables and Price

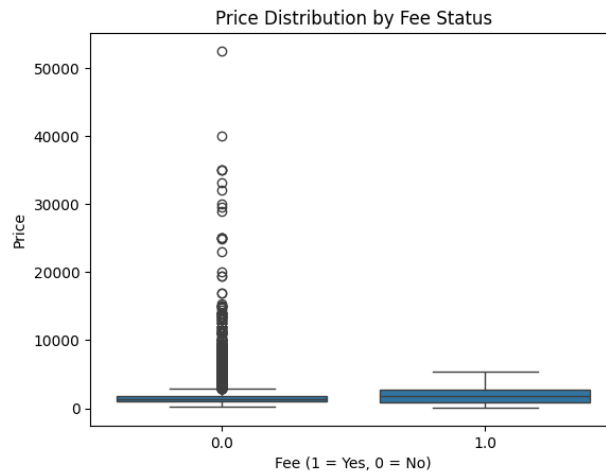


Figure 5.6: Comparison of Monthly Rental Prices for Apartments With and Without Upfront Fees

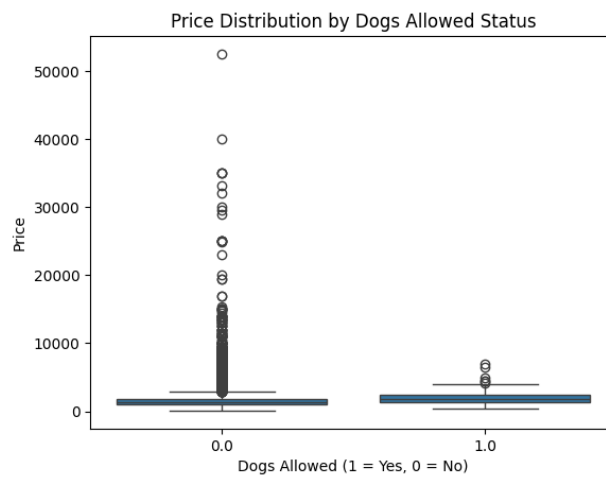


Figure 5.7: Comparison of Monthly Rental Prices Based on Dog Allowance Policies

The boxplots in Figures 5.6 and 5.7 show the distribution of price for the categorical variable representing whether the apartment listing has an initial fee (Figure 5.6), and the categorical variable on whether the apartment allows dogs (Figure 5.7). This allows us to

compare the price of apartments depending on whether it has a fee and whether it allows dogs. The mean price is higher for listings that had no fees and allow dogs, although for both variables, there are many high outliers for apartments without fees and apartments that do not allow dogs. The categorical variable `fee_Yes` and `pets_allowed_dogs` had statistically significant differences in mean rent prices. It can be assumed that expensive apartments have less fees and do not allow dogs.

The variable for `latitude`'s statistical summary is as follows: the mean was 3.15e04, standard deviation was 7.02e06, minimum was 1.96e01, maximum was 1.57e09, and median was 3.72e01. The `longitude` variable's statistical summary is as follows: the mean was -91.56, standard deviation was 15.82, minimum was -159.37, maximum was -68.77, and median was -84.56. `latitude` and `longitude` had a slight correlation with price, with `latitude` -0.00 and `longitude` -0.11. The `fee` variable's statistical summary is as follows: there were four unique values, the most frequent was No, with 99556 apartments that had no fees. The `pets_allowed` variable's statistical summary is as follows: there were five unique values, the most frequent was "Cats, Dogs" and 97,781 apartments allowed both cats and dogs.

State and Price

The bar graph in Figure 5.8 (Figure 5.8) shows average apartment listing prices for all states in the United States. It is in descending order, with the highest average listing price on the left. It shows us the ranking of states' average apartment listing prices. This graph shows that the state with the highest mean price was Hawaii (\$2,729.19), then New York (\$2,553.01), then California (\$2,461.16). The lowest apartment listing average is in Wyoming, then New Mexico, then West Virginia. The states with the highest average apartment listings were also the states with the highest correlation to the price variable.

The map of the United States in Figure 5.9 (Figure 5.9) gives us a spatial and geographic view of the average apartment prices in the United States, per state, which allows us to analyze how location plays a role in apartment price trends.

This map shows that the highest average apartment prices are mainly in states that are on the coasts of the United States or in Hawaii, and the cheapest average prices are in states in the Central region. The medium priced listings are mainly from states in between the high and low price states.

Out of these three states, the correlation to price was highest for California (0.35) then Hawaii (0.02), then New York (0.09). The difference in mean price is statistically significant from a t-test conducted comparing the mean prices ($p\text{-value} = 0.0$) for listings in these three states compared to the rest of the dataset.

5 Exploratory Data Analysis - California

Correlation Heatmap

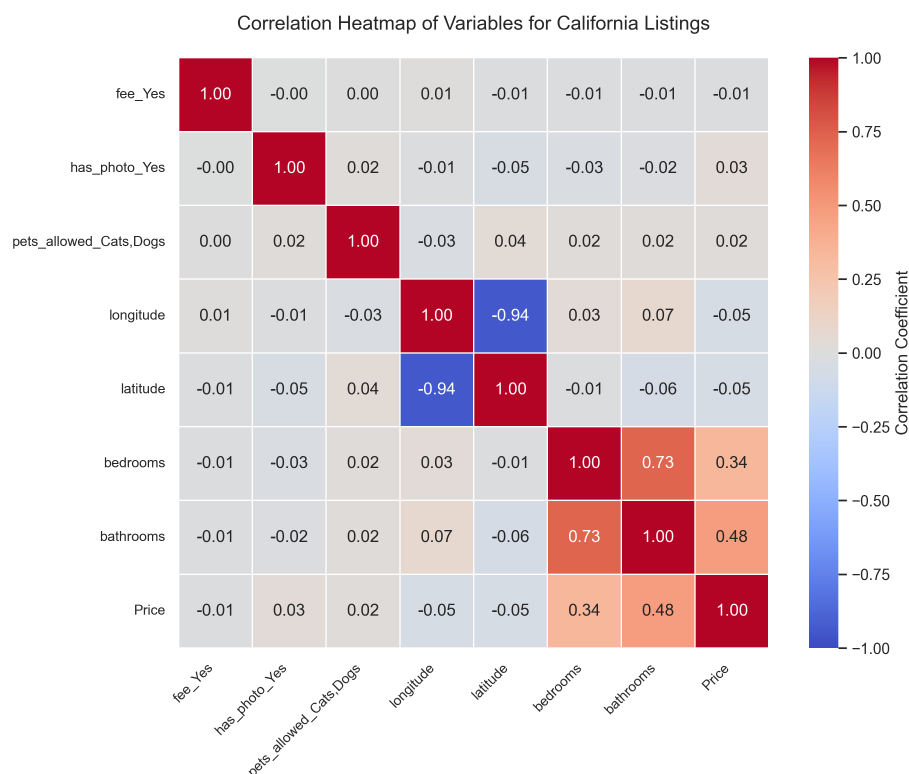


Figure 5.10: Correlation Heatmap of Property Features for California Listings

The correlation heatmap in Figure 5.10 (Figure 5.10) shows that variables with the

highest correlation to price from the ones chosen for this analysis are **bathrooms** (0.48) and **bedrooms** (0.34). The other variables did not have high correlation, but can still have a nonlinear or complex interaction with price. Other high correlations were between the number of bedrooms and bathrooms (0.73) and between longitude and latitude (-0.94).

Price

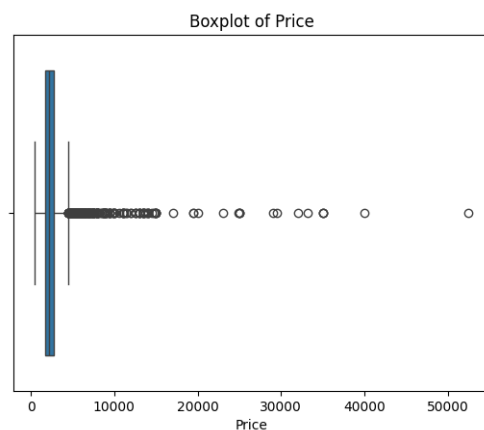


Figure 5.11: Distribution of Monthly Rental Prices for California Apartments

There were many high outliers for price (Figure 5.11), and the median price was \$2,195.00 while the mean price was \$2,463.47 with standard deviation \$1,709.16. The minimum, \$493, was for a listing in Newman that had 1 bathroom, two bedrooms, Cats and Dogs allowed. The maximum, \$52,500, was for a listing in Barstow that had one bathroom and no bedrooms, where Cats and Dogs were allowed. This was the same maximum listing in the national unfiltered dataset as well.

The mean price per square foot in California listings was \$2.80/sqft, which is higher than the national average price per square foot. The median was \$2.60/sqft, which is also higher. The minimum was \$0.16/sqft while maximum was \$37.02/sqft. The minimum is higher than the national minimum, while maximum is the same. The standard deviation was \$1.20/sqft, which is also higher than the national price per square foot standard deviation.

Bathrooms and Price

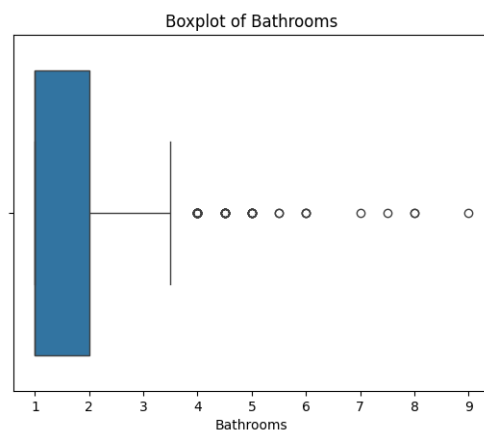


Figure 5.12: Distribution of the Number of Bathrooms in California Apartments

There were 29 high outliers for the bathroom variable (Figure 5.12). The mean number of bathrooms was 1.45 with a standard deviation of 0.61, and the median was 1. The P-value was statistically significant for all bathroom comparisons (bathrooms 0 vs 1, 1 vs 2, 2 vs 3, 3 vs 4).

Bedrooms and Price

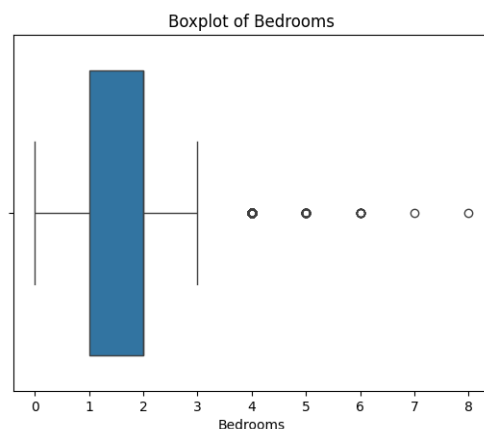


Figure 5.13: Distribution of the Number of Bedrooms in California Apartments

There were 49 high outliers for the `bedrooms` variable (Figure 5.13). The mean `bedrooms` was 1.67, with a standard deviation of 0.76, and the median was 2. There was a statistically

significant difference in mean rent prices between bedroom changes by 1 (for rooms 0 to 4). The maximum number of bedrooms is eight, and this is for a listing at 40 North Harbor Boulevard, San Pedro, CA 90731 [7]. The price for this listing is \$2,700/month.

Other Variables and Price

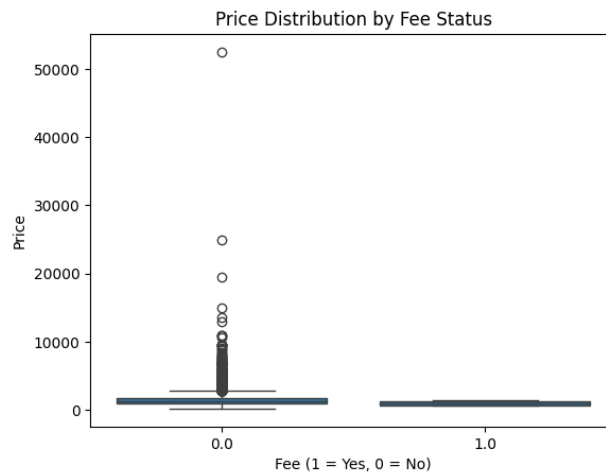


Figure 5.14: Comparison of Rental Price Distributions for California Apartments With and Without Upfront Fees

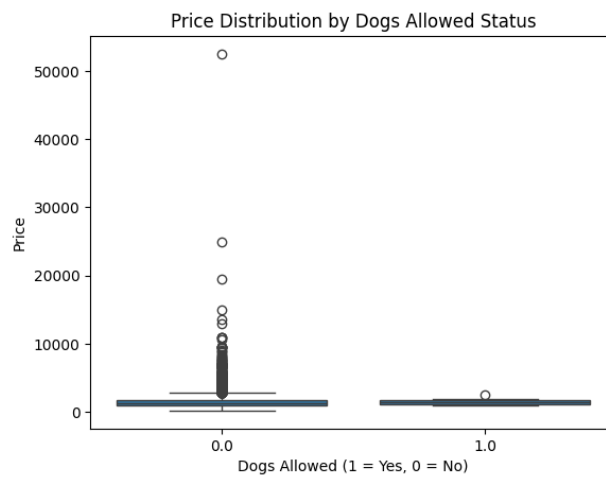


Figure 5.15: Comparison of Rental Price Distributions for California Apartments Based on Dog Allowance

The variable for `latitude`'s statistical summary is as follows: the mean was -118.57,

standard deviation was 1.68, minimum was 32.56, maximum was 40.93, and median was 34.03. The `longitude` variable’s statistical summary is as follows: the mean was -118.57, standard deviation was 1.55, minimum was -124.23, maximum was -115.57, and median was -118.26.

`Latitude` and `longitude` had a slight correlation with price per square foot, with `latitude` -0.03 and `longitude` 0.15. `Longitude` had a stronger correlation to price per square foot than `latitude`.

The `fee` variable’s statistical summary is as follows: there were 2 unique values, the most frequent was No, with 10,307 apartments that had no fees. The `pets_allowed` variable’s statistical summary is as follows: there were 3 unique values, the most frequent was “Cats, Dogs” and 10,029 apartments allowed both cats and dogs.

The categorical variable `fee_Yes` and `pets_allowed_dogs` had statistically significant differences in mean rent prices. The mean price is higher for listings that had no fees (Figure 5.14) and allow dogs (Figure 5.15). However, for both variables, there are many high outliers for apartments without fees and apartments that do not allow dogs.

State and Price

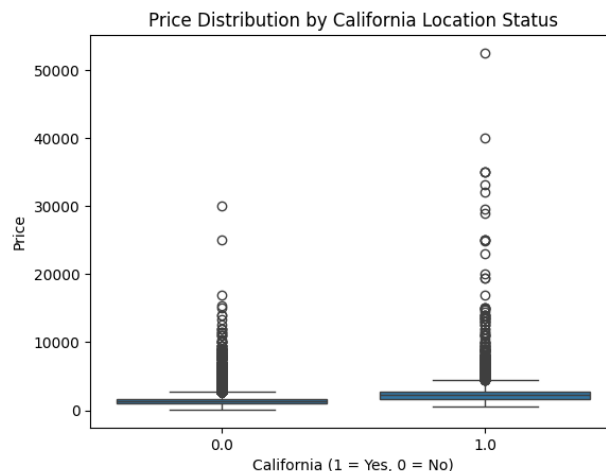


Figure 5.16: Comparison of Rental Price Distributions for California Apartments and Non-California Apartments

The mean price for apartments in California was \$2,463.77, while the mean price of apartments in the dataset that were not listed in California was \$1,418.97.

The boxplot in Figure 5.16 (Figure 5.16) shows a comparison of apartment monthly rental price between listings located in California, and in other states. This can visualize the hypothesis testing done to compare the prices of apartments not in California and in California.

It shows that listings in California have many high outliers that are above the prices of other listings in other states of the United States. In addition, all quartiles are higher than the listings in other states. Thus, we can assume that on average, California apartments are higher than most other states' listings. When hypothesis testing was done for comparison, it shows that there is a statistically significant difference in price for apartments not in California and in California. The mean listing price was \$2,461.17 for apartments in California, and \$1,418.96 for apartments in other states.

The map of California apartment listings in Figure 5.17 (Figure 5.17) shows that most of the listings in the low price category seem to be away from the coast, and as listings get closer to the coast, the price seems to increase. This trend is visible for both Northern and Southern California. In the predictive modeling section, I will analyze this trend further using predictive modeling with a "distance to the nearest beach" variable.

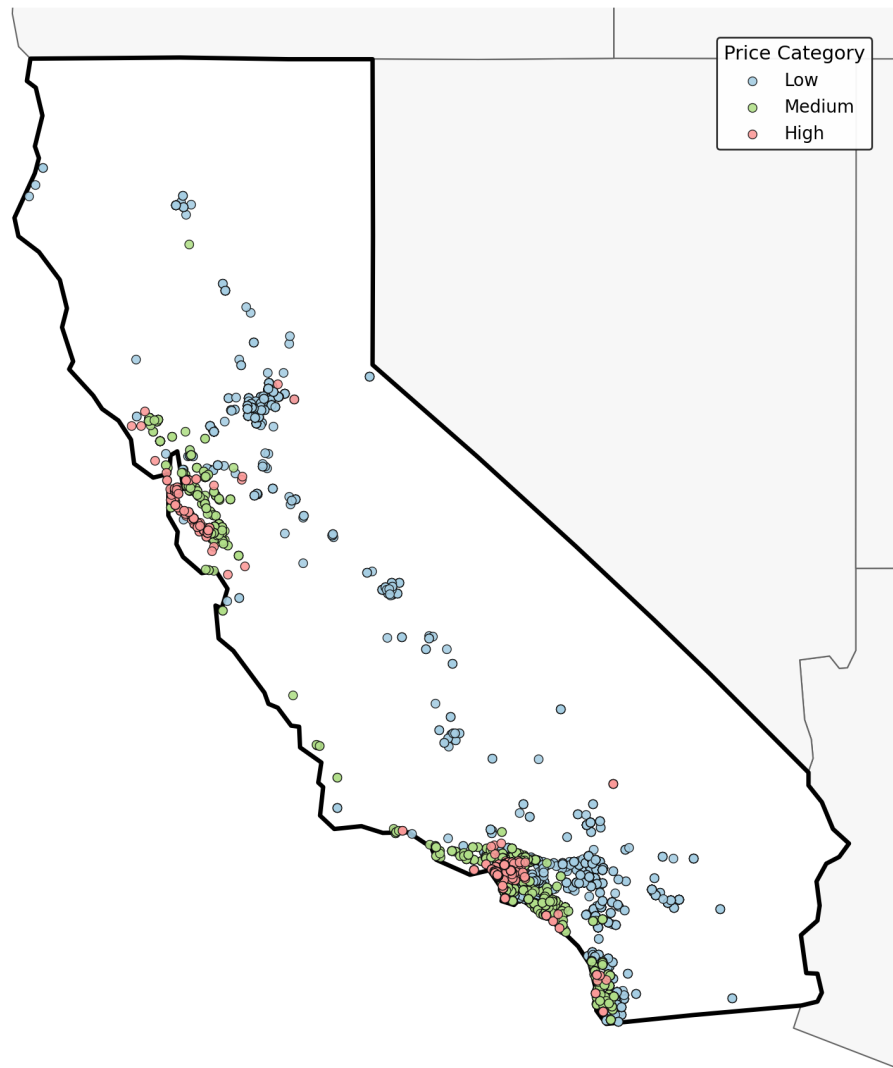


Figure 5.17: Geographic Distribution of Mean California Apartment Rental Prices by Price Tier

CHAPTER 6

Modeling

Predictive Modeling - National Listings

Results

Model	Number of Feature Variables	RMSE	R^2
Random Forest	4	242.61	0.79
Random Forest	8	244.55	0.79
XGBoost	8	270.2	0.74
XGBoost	4	271.33	0.74
Gradient Boosting	8	340	0.59
Gradient Boosting	4	340.24	0.58
Multivariate Linear Reg	8	454.25	0.26
Multivariate Linear Reg	4	499.63	0.11

Table 6.1: Model Performance Excluding Price Outliers for U.S. Listings

In these tables, the models are in the order of lowest RMSE and highest R^2 . The four feature variables are: `longitude`, `latitude`, `bedrooms`, and `bathrooms`. The eight feature variables are: `fee_Yes`, `has_photo_Yes`, `pets_allowed_Dogs`, `state_CA`, `longitude`, `latitude`, `bedrooms`, and `bathrooms`.

When price outliers were excluded, Random Forest performed the best for predicting price from eight selected variables as well as just four numeric selected variables. It did slightly better with just the four numeric variables with a smaller root mean squared error (RMSE) (242.61 vs 244.55), but equivalent R^2 (0.79). Thus, it explained 79% of the variance in price

Model	Number of Feature Variables	RMSE	R^2
Random Forest	8	453.3	0.75
Random Forest	4	456.49	0.74
XGBoost	4	501.98	0.69
XGBoost	8	512.18	0.68
Gradient Boosting	8	552.89	0.62
Gradient Boosting	4	555.06	0.62
Multivariate Linear Reg	8	765.79	0.28
Multivariate Linear Reg	4	838.88	0.14

Table 6.2: Model Performance Including Price Outliers for U.S. Listings

through the models, but the model with just four numeric variables had a smaller root mean squared error. An RMSE of 456.49 indicates that the typical error size in the predicted monthly rental fee is \$456.49, highlighting that the best model here still exhibits relatively large prediction errors in general. For all models, modeling with just the four numeric variables had better metrics than using all eight. For Multivariate Linear Regression, the model with price outliers performed better than the model without price outliers. For all other models, (Random Forest, XGBoost, and Gradient Boost), the model without price outliers performed better than the model with price outliers. For feature importance from Random Forest, XGBoost, and Gradient Boost, location in California, number of `bathrooms`, and `longitude` were consistently in the top three. The order for California and `longitude` switched for XGBoost compared to the other two models: for XGBoost, the top three in order were California, `bathrooms`, `longitude` but for both Random Forest and Gradient Boosting, the top three in order were `longitude`, `bathrooms`, and California.

Kernel Regression

A C function was made and called by Python for both Kernel Regression and GAM (only for the Kernel smoothing). The outcomes were the same for Kernel Regression and GAM for only four numeric variables, as both were based on additive Kernel Smoothing methods.

Models were repeated with data without price outliers. GAM and Kernel Regression were both done on 5000 sampled data. When including the categorical binary variables and using linear fitting (for eight variables), this slightly affected the R squared or RMSE. For GAM on data without price outliers, when using eight variables, the R^2 was 0.44 and RMSE was 384.08, When using four numeric variables, the R^2 was 0.43 and RMSE was 388.00, so it showed a slight improvement. For GAM on data with price outliers, when using eight variables, the R^2 was 0.65 and RMSE was 574.48. Kernel Regression for the `price` variable still exhibited relatively large prediction errors. When using four numeric variables, the R^2 was 0.65 and RMSE was 577.54, which shows a slight increase in RMSE.

For just numeric variables, the GAM and Kernel Regression equation is:

$$\text{price} = f_1(\text{longitude}) + f_2(\text{latitude}) + f_3(\text{bathrooms}) + f_4(\text{bedrooms}) + \varepsilon.$$

For all 8 variables, the GAM equation is:

$$\begin{aligned} \text{price} = & f_1(\text{longitude}) + f_2(\text{latitude}) + f_3(\text{bathrooms}) + f_4(\text{bedrooms}) \\ & + \beta_5(\text{state_CA}) + \beta_6(\text{fee_Yes}) + \beta_7(\text{has_photo_Yes}) + \beta_8(\text{pets_allowed_Dogs}) + \varepsilon. \end{aligned}$$

Model Performance Before Price Outlier Removal

Before removing price outliers, Kernel Regression was done on just the numeric variables where the order of Kernel Smoothing and equation was:

$$\text{price} = f_1(\text{bathrooms}) + f_2(\text{bedrooms}) + f_3(\text{longitude}) + f_4(\text{latitude}) + \varepsilon.$$

It explained 82% of the variance of price with \$408.81 root mean squared error, which suggests that it is the best performing model out of all models conducted.

Model Performance After Changing Response Variable to Price per Square Foot

In an attempt to normalize the target metric across varying property sizes, the response variable was changed from the total monthly price to the price per square foot (`price_per_square_foot`). This structural pivot directly alters the interpretation and scale of the baseline error metrics.

When transitioning to modeling for `price_per_square_foot`, the apparent drop in the Root Mean Squared Error to values as low as 0.43 (such as seen in the Random Forest models) is entirely a consequence of the new measurement scale: an RMSE of 0.43 indicates that the typical prediction error is \$0.43 per square foot. For a standard 1,000-square-foot apartment listing, this relative metric scales back up to a total absolute monthly prediction error of approximately \$430.00. Thus, while changing the response variable successfully levels the prices and compares properties fairly regardless of physical footprint, the underlying prediction errors remain structurally substantial.

United States Apartments, Including Price Outliers

Model	Number of Feature Variables	RMSE	R^2
Random Forest	4	0.43	0.74
Random Forest	8	0.43	0.74
XGBoost	4	0.46	0.71
XGBoost	8	0.46	0.71
Gradient Boosting	4	0.54	0.6
Gradient Boosting	8	0.54	0.6
Multivariate Linear Reg	8	0.71	0.3
Multivariate Linear Reg	4	0.8	0.1

Table 6.3: Model Performance Including Price per Square Foot Outliers for U.S. Listings

When using all eight variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.71 with an R-squared of 0.30. A RMSE of 0.71 means that the prediction error is \$0.71 per square foot per month. Random Forest had an RMSE of 0.43 and an R-squared of 0.74, producing the best results out of the four tested in these conditions. The feature importance for Random Forest was (from most important to least): `longitude`, `latitude`, `state_CA`, `bedrooms`, `bathrooms`, `has_photo_Yes`. All other variables had minimal importance to the model. XGBoost had a RMSE of 0.46 and an R-squared of 0.71. The feature importance for XGBoost was (from most important to least): `state_CA`, `longitude`,

`latitude`, `bedrooms`, and `bathrooms`. All other variables had minimal importance for the model. Finally, Gradient Boost had a RMSE of 0.54 and an R-squared of 0.60. The feature importance for Gradient Boost was (from most important to least): `longitude`, `latitude`, `state_CA`, `bedrooms`, and `bathrooms`. All other variables had minimal importance for the model.

When using just the four numeric variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.80 with an R-squared of 0.10. Random Forest had an RMSE of 0.43 and an R-squared of 0.74, and this was the best performing model in this comparison. The feature importance for Random Forest was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`. XGBoost had a RMSE of 0.46 and an R-squared of 0.71. The feature importance for XGBoost was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`. Finally, Gradient Boost had a RMSE of 0.54 and an R-squared of 0.60. The feature importance for Gradient Boost was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`.

United States Apartments, Excluding Price Outliers

Model	Number of Feature Variables	RMSE	R^2
Random Forest	4	0.29	0.76
Random Forest	8	0.3	0.75
XGBoost	8	0.31	0.72
XGBoost	8	0.32	0.72
Gradient Boosting	8	0.38	0.58
Gradient Boosting	8	0.38	0.58
Multivariate Linear Reg	8	0.51	0.26
Multivariate Linear Reg	8	0.56	0.11

Table 6.4: Model Performance Excluding Price per Square Foot Outliers for U.S. Listings

When using all eight variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.51 with an R-squared of 0.26. Random Forest had an RMSE

of 0.30 and an R-squared of 0.75, which is the best result out of these models. The feature importance for Random Forest was (from most important to least): `longitude`, `latitude`, `bedrooms`, `bathrooms`, `has_photo_Yes`. All other variables had minimal importance to the model. XGBoost had a RMSE of 0.32 and an R-squared of 0.72. The feature importance for XGBoost was (from most important to least): `longitude`, `latitude`, `bedrooms`, `state_CA`, `bathrooms`, `has_photo_Yes`, `fee_Yes`, and `pets_allowed_Dogs`. Finally, Gradient Boost had a RMSE of 0.38 and an R-squared of 0.58. The feature importance for Gradient Boost was (from most important to least): `longitude`, `latitude`, and `bedrooms`. All other variables had minimal importance for the model.

When using just the four numeric variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.56 with an R-squared of 0.11. Random Forest had an RMSE of 0.29 and an R-squared of 0.76, producing the best results out of the four tested here. The feature importance for Random Forest was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`. XGBoost had a RMSE of 0.31 and an R-squared of 0.72. The feature importance for XGBoost was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`. Finally, Gradient Boost had a RMSE of 0.38 and an R-squared of 0.58. The feature importance for Gradient Boost was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`.

California Apartments, Including Price Outliers

When using all seven variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.95 with an R-squared of 0.33. Random Forest had an RMSE of 0.72 and an R-squared of 0.61. The feature importance for Random Forest was (from most important to least): `longitude`, `latitude`, `bedrooms`, `bathrooms`, `has_photo_Yes`, `pets_allowed_Dogs`, `fee_Yes`. All other variables had minimal importance to the model. XGBoost had a RMSE of 0.71 and an R-squared of 0.63, which was the strongest model here. The feature importance for XGBoost was (from most important to least): `longitude`, `latitude`, `bathrooms`, `bedrooms`, `has_photo_Yes`, `pets_allowed_Dogs`, `fee_Yes`. All other

Model	Number of Feature Variables	RMSE	R^2
XGBoost	4	0.71	0.63
XGBoost	7	0.71	0.63
Random Forest	4	0.72	0.62
Random Forest	7	0.72	0.61
Gradient Boosting	4	0.75	0.59
Gradient Boosting	7	0.75	0.59
Multivariate Linear Reg	7	0.95	0.33
Multivariate Linear Reg	4	0.96	0.33

Table 6.5: Model Performance Including Price per Square Foot Outliers for CA Listings

variables had minimal importance for the model. Finally, Gradient Boost had a RMSE of 0.75 and an R-squared of 0.59. The feature importance for Gradient Boost was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`. All other variables had minimal importance for the model.

When using just the four numeric variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.96 with an R-squared of 0.33. Random Forest had an RMSE of 0.72 and an R-squared of 0.62. The feature importance for Random Forest was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`. XGBoost had a RMSE of 0.71 and an R-squared of 0.63, performing the best in this set. The feature importance for XGBoost was (from most important to least): `longitude`, `latitude`, `bathrooms`, and `bathrooms`. Finally, Gradient Boost had a RMSE of 0.75 and an R-squared of 0.59. The feature importance for Gradient Boost was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`.

California Apartments, Excluding Price Outliers

When using all seven variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.70 with an R-squared of 0.44. Random Forest had an RMSE

Model	Number of Feature Variables	RMSE	R^2
XGBoost	4	0.48	0.73
XGBoost	7	0.49	0.73
Random Forest	4	0.5	0.72
Random Forest	7	0.51	0.71
Gradient Boosting	4	0.54	0.67
Gradient Boosting	7	0.54	0.67
Multivariate Linear Reg	4	0.7	0.44
Multivariate Linear Reg	7	0.7	0.44

Table 6.6: Model Performance Excluding Price per Square Foot Outliers for CA Listings

of 0.51 and an R-squared of 0.71. The feature importance for Random Forest was (from most important to least): `longitude`, `latitude`, `bedrooms`, `bathrooms`, `has_photo_Yes`. All other variables had minimal importance to the model. XGBoost had a RMSE of 0.49 and an R-squared of 0.73, which is the best model in this set. The feature importance for XGBoost was (from most important to least): `latitude`, `longitude`, `bathrooms`, `bedrooms`, `has_photo_Yes`, `pets_allowed_Dogs`, and `fee_Yes`. Finally, Gradient Boost had a RMSE of 0.54 and an R-squared of 0.67. The feature importance for Gradient Boost was (from most important to least): `longitude`, `latitude`, `bedrooms`, `bathrooms`. All other variables had minimal importance for the model.

When using just the four numeric variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.70 with an R-squared of 0.44. Random Forest had an RMSE of 0.50 and an R-squared of 0.72. The feature importance for Random Forest was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`. XGBoost had a RMSE of 0.48 and an R-squared of 0.73, performing the best out of this set. The feature importance for XGBoost was (from most important to least): `latitude`, `longitude`, `bathrooms`, and `bedrooms`. Finally, Gradient Boost had a RMSE of 0.54 and an R-squared of 0.67. The feature importance for Gradient Boost was (from most important to least): `longitude`, `latitude`, `bedrooms`, and `bathrooms`.

Model Performance After Adding a Distance from Nearest Beach Variable for California Apartments

A new predictor variable was made for predicting the price of California apartments specifically, and this was the distance to nearest beach in kilometers (`dist_to_nearest_beach_km`). These were the results when adding this field in, and I ran these models to compare the performance to the California apartment price modeling above.

Including Price per Square Foot Outliers

Model	Number of Feature Variables	RMSE	R^2
XGBoost	5	0.71	0.63
XGBoost	8	0.72	0.62
Random Forest	5	0.72	0.62
Random Forest	8	0.73	0.61
Gradient Boosting	8	0.75	0.59
Gradient Boosting	5	0.75	0.59
Multivariate Linear Reg	8	0.94	0.36
Multivariate Linear Reg	5	0.94	0.36

Table 6.7: Model Performance Including Price per Square Foot Outliers for CA Listings with Distance to Nearest Beach

When using all eight variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.94 with an R-squared of 0.36. Random Forest had an RMSE of 0.73 and an R-squared of 0.61. The feature importance for Random Forest was (from most important to least): `latitude`, `dist_to_nearest_beach_km`, `longitude`, `bedrooms`, `bathrooms`, and `has_photo_Yes`. All other variables had minimal importance to the model. XGBoost had a RMSE of 0.72 and an R-squared of 0.62, which was the best in this set. The feature importance for XGBoost was (from most important to least): `dist_to_nearest_beach_km`, `latitude`, `longitude`, `bathrooms`, `bedrooms`, `has_photo_Yes`,

`pets_allowed_Dogs`, and `fee_Yes`. Finally, Gradient Boost had a RMSE of 0.75 and an R-squared of 0.59. The feature importance for Gradient Boost was (from most important to least): `latitude`, `dist_to_nearest_beach_km`, `longitude`, `bedrooms`, and `bathrooms`. All other variables had minimal importance for the model.

When using just the five numeric variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.94 with an R-squared of 0.36. Random Forest had an RMSE of 0.72 and an R-squared of 0.62. The feature importance for Random Forest was (from most important to least): `latitude`, `dist_to_nearest_beach_km`, `longitude`, `bedrooms`, and `bathrooms`. XGBoost had a RMSE of 0.71 and an R-squared of 0.63, producing the best results out of the four tested. The feature importance for XGBoost was (from most important to least): `dist_to_nearest_beach_km`, `latitude`, `longitude`, `bathrooms`, and `bedrooms`. Finally, Gradient Boost had a RMSE of 0.75 and an R-squared of 0.59. The feature importance for Gradient Boost was (from most important to least): `latitude`, `dist_to_nearest_beach_km`, `longitude`, `bedrooms`, and `bathrooms`.

Excluding Price per Square Foot Outliers

Model	Number of Feature Variables	RMSE	R^2
XGBoost	8	0.48	0.74
XGBoost	5	0.49	0.73
Random Forest	5	0.5	0.72
Random Forest	8	0.5	0.71
Gradient Boosting	8	0.53	0.67
Gradient Boosting	5	0.54	0.67
Multivariate Linear Reg	5	0.69	0.46
Multivariate Linear Reg	8	0.69	0.46

Table 6.8: Model Performance Excluding Price per Square Foot Outliers for CA Listings with Distance to Nearest Beach

When using all eight variables, here were the results. Linear Regression had a Root

Mean Squared Error (RMSE) of 0.69 with an R-squared of 0.46. Random Forest had an RMSE of 0.50 and an R-squared of 0.71. The feature importance for Random Forest was (from most important to least): `latitude`, `dist_to_nearest_beach_km`, `longitude`, `bathrooms`, `bedrooms`, and `has_photo_Yes`. All other variables had minimal importance to the model. XGBoost had a RMSE of 0.48 and an R-squared of 0.74, once again producing the best results. The feature importance for XGBoost was (from most important to least): `dist_to_nearest_beach_km`, `latitude`, `longitude`, `bathrooms`, `bedrooms`, `has_photo_Yes`, `pets_allowed_Dogs`, and `fee_Yes`. Finally, Gradient Boost had a RMSE of 0.53 and an R-squared of 0.67. The feature importance for Gradient Boost was (from most important to least): `latitude`, `dist_to_nearest_beach_km`, `longitude`, `bathrooms`, and `bedrooms`. All other variables had minimal importance for the model.

When using just the five numeric variables, here were the results. Linear Regression had a Root Mean Squared Error (RMSE) of 0.69 with an R-squared of 0.36. Random Forest had an RMSE of 0.50 and an R-squared of 0.72. The feature importance for Random Forest was (from most important to least): `latitude`, `dist_to_nearest_beach_km`, `longitude`, `bathrooms`, and `bedrooms`. XGBoost had a RMSE of 0.49 and an R-squared of 0.73, consistently producing the best model in these sets. The feature importance for XGBoost was (from most important to least): `dist_to_nearest_beach_km`, `latitude`, `longitude`, `bathrooms`, and `bedrooms`. Finally, Gradient Boost had a RMSE of 0.54 and an R-squared of 0.67. The feature importance for Gradient Boost was (from most important to least): `latitude`, `dist_to_nearest_beach_km`, `longitude`, `bathrooms`, and `bedrooms`.

Overall, adding distance to the nearest beach only made the RMSE worsen by 0.01 but had the same R-squared as the best performing predictive model for California apartments (XGBoost without price outliers and with four numeric feature variables). This is because passing `latitude` and `longitude` to the model already implicitly adds the distance to the nearest beach, and so we are adding redundant information that the model has already captured from coordinates, which may have caused a worse generalization. In addition, distance to the beach matters more in Southern California rather than Northern California, due to the sunny weather and high value of being close to the beach. These external factors

need to be taken into account to improve the model as well.

More Models after Feature Engineering

I also performed two more experiments, which were removing the `longitude` and `latitude` variables and just keeping the new distance to the nearest beach variable. The second experiment is to do a clean categorical cut with binary flags related to being closer to the beachfront. As described in the feature engineering section of the Data chapter, I added "`coastal_lat_interaction`" to differentiate the distance to the beach for Northern and Southern California, and "`is_waterfront`" for a binary variable for apartments with one kilometer (km) distance to the beach.

Here were the results: When the features used were `longitude`, `latitude`, `bedrooms`, `bathrooms`, `is_waterfront`, `coastal_lat_interaction`, the RMSE was 0.49. The R-squared was 0.7318.

When the features used were `bedrooms`, `bathrooms`, `dist_to_nearest_beach_km`, `is_waterfront`, `coastal_lat_interaction`, the RMSE was 0.5301 and the R-squared was 0.6805.

The RMSE and R-squared worsened after these two experiments, which shows that `latitude` and `longitude` is a better predictor than just using just distance to the nearest beach, as it has information on the complete mapping of the listing, not just a linear path to the beach. `latitude` and `longitude` have more than just data on the location, it also may include information such as zip code wealth, local job markets, and school districts, that all make the model better at predicting price.

Best Models and Residual Analysis

For national apartments, the best predictive model from my analysis was Random Forest using four numeric features: `latitude`, `longitude`, `bathrooms`, and `bedrooms`. The target variable was price per square foot, with outliers removed with the IQR method. This model had a RMSE of 0.29 and a R-squared of 0.76.

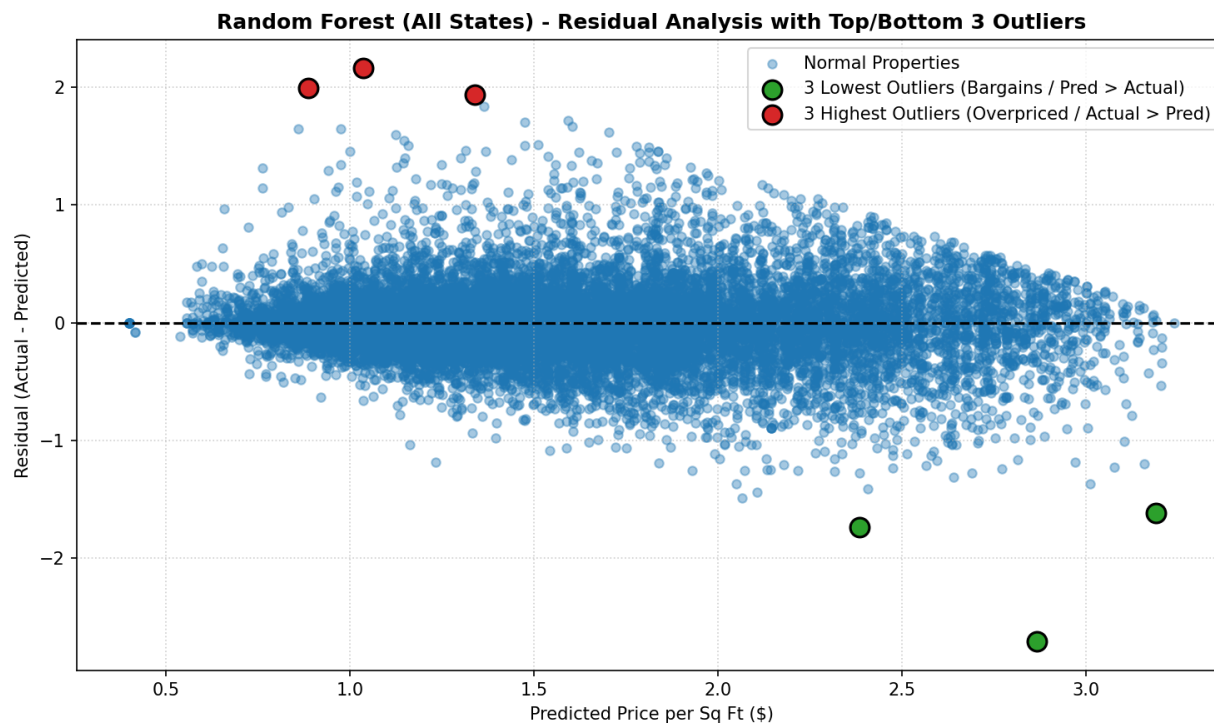


Figure 6.1: Model Residuals: Best Model for U.S. Apartments

As the residual graph (Figure 6.1) has a balanced scatter above and below the horizontal line at zero, this means that the model has successfully captured all the meaningful predictive signals from the features. The remaining errors seem to be complete random noise.

For California apartments, the best model was XGBoost using four same numeric features: `latitude`, `longitude`, `bathrooms`, and `bedrooms`. The target variable was also price per square foot, with outliers removed with the IQR method. This model had a RMSE of 0.48 and a R-squared of 0.73.

The model's residual graph (Figure 6.2) also has a balanced scatter above and below the horizontal line at zero, and shows that the model has successfully captured all the meaningful predictive signals from the features. The remaining errors seem to be complete random noise.

Similarities from both models show that spatial features are heavily predictive of price per square foot of the apartment, and perform better in prediction compared to categorical features of listings such as whether the apartment has a fee before purchase, whether the listing has a photo, whether dogs are allowed, and whether the apartment is in California.

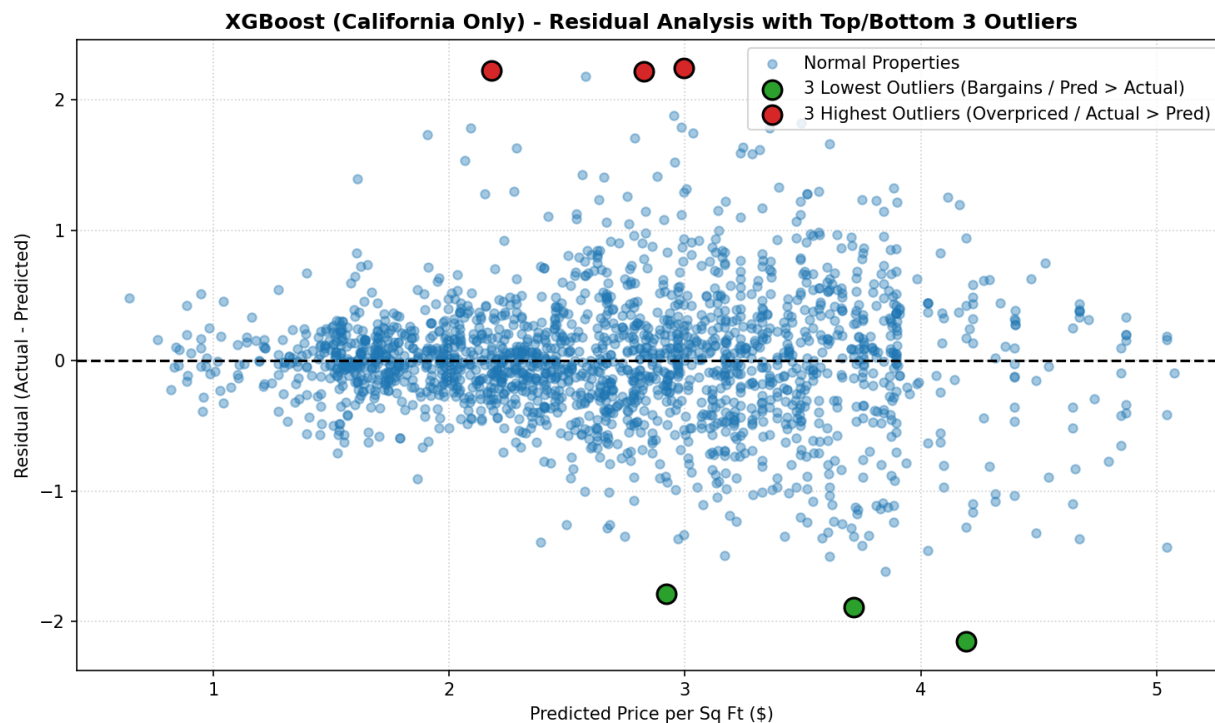


Figure 6.2: Model Residuals: Best Model for CA Apartments

Price Analysis from the Best Models

When analyzing the residuals from the best national apartments model, I found the top three underpriced apartments that had an actual price less than predicted. In addition, I also found top three overpriced apartments that had actual prices more than what was predicted.

For the national apartment model (Random Forest), the top three underpriced listings were in Los Angeles, CA (-\$2.71/sqft), Winooski, VT (-\$1.73/sqft), and Philadelphia, PA (-\$1.62/sqft). It was surprising to see that the most underpriced listing was in Los Angeles, CA. Winooski, VT is near a college town, and Philadelphia also has high student populations, so these bargain listings may be due to student housing trends and off-season trends.

The top three overpriced listings were in Little Rock, AR (+\$2.16/sqft), Norfolk, VA (+\$1.99/sqft), and Shorewood, WI (+\$1.93/sqft). This was also surprising, as none of these cities are traditionally considered having a high cost of living, such as New York or San

Francisco. Rental spikes in Norfolk and Shorewood may come from a sudden high demand in smaller markets, but these are also external factors that this model did not know about.

This may suggest that this model is capturing macroeconomic shifts or certain local pockets are experiencing rapid rent inflation.

For the California apartment model (XGBoost), the top three bargain listings were in Los Angeles (-\$2.15/sqft and -\$1.89/sqft), and Long Beach (-\$1.79/sqft). Los Angeles shows up again in this model, so this underpriced listing is a bargain price from both models' residuals.

The top three overpriced listings were in Marina del Ray (+\$2.24/sqft), San Diego (+\$2.22/sqft), and Beverly Hills (+\$2.22/sqft). This result makes intuitive sense, as Marina del Ray and Beverly Hills are very expensive neighborhoods in Southern California, and may have a luxury tax that this model does not take into account.

CHAPTER 7

Conclusion

Interpretation of Findings

The key takeaway from this analysis and modeling is that geographic location is overwhelmingly the most powerful driver of apartment rent prices across the United States, and this is especially true within California. We can see this clearly through the heavy predictive weight that latitude and longitude hold in the machine learning models, as well as the high correlation between an apartment being in California and having a higher monthly rent. This spatial trend is further highlighted when looking at state-level averages. Hawaii, New York, and California had the highest mean rental costs in the dataset, which suggests a clear pricing premium tied to coastal proximity. In contrast, the cheapest average rent prices were found in states far from coasts, such as Wyoming, New Mexico, and West Virginia. On average, listings in coastal regions consistently were more expensive than those in the central regions of the country.

Beyond geographic location, structural variables also played a large role in driving rent variations. Our statistical analysis confirmed that rent increases alongside layout features, as a higher number of bedrooms and bathrooms directly corresponds to a higher monthly price. While listing-level administrative details and rules, such as whether a property charges an upfront fee or allows dogs, did show statistically significant impacts on average costs, they ultimately carried much less weight than location and layout features. When it came to predictive modeling, structural layout and spatial coordinates had far more predictive power than any of these categorical listing amenities.

Limitations

Generalizability

Several data and structural limitations should be taken into account when looking at the generalizability of these findings. First, the dataset suffers from a geographic imbalance and sampling bias, heavily over-representing listings from Texas, California, and Virginia. This regional skewness makes it difficult for a global model to accurately capture localized pricing trends in under-represented states. Second, our engineered feature for coastal proximity introduced spatial representation errors. Because we relied on a discrete list of 28 California beach coordinates to measure distance, the model missed the complex, jagged nature of the actual coastline. As a result, properties that are technically right on the water but located between our designated data points were incorrectly calculated as being deep inland, creating conflicting signals for the learning algorithm.

Additionally, creating derivative spatial variables, such as multiplying beach distance by latitude to capture a geographic interaction, caused heavy multi-collinearity. In tree-based models like XGBoost, feature engineering spatial variables diluted the baseline coordinates and degraded the model’s ability to generalize to new data, which ultimately lowered performance. Finally, while extreme price outliers were truncated using an Interquartile Range (IQR) mask, structural outliers within the inputs, such as properties with anomalous bedroom or bathroom counts, were left in the data.

Model Residual Errors

A primary limitation of this study lies in the absolute magnitude of the predictive errors across both the national and regional modeling workflows.

These substantial residual errors demonstrate that while structural features (**bedrooms**, **bathrooms**) and macroscopic spatial coordinates (**latitude**, **longitude**) capture broad geographic trends and property layouts, they introduce a notable degree of omitted variable bias. Real estate valuations are intrinsically bound to highly granular, unobserved qualitative

features that were absent from the scraped web-listing data. Factors such as interior luxury finishes, property age, community amenities (e.g., fitness centers, secure parking), micro-neighborhood safety metrics, localized school district boundaries, and short-term macroeconomic supply shocks heavily impact pricing premiums. Consequently, as the estimators are blind to these hidden parameters, the residual variance remains structurally high, and the models from this thesis should be interpreted as regional baseline evaluators.

Future Work

To build on these results, future research should focus on a few key improvements. As raw spatial coordinates cannot capture the actual interior condition of a property, a crucial next step is to use the "description" variable, which is an unstructured text field. By parsing raw text descriptions for descriptive keywords like "luxury," "renovated," or "in-unit laundry," we can extract structured features to capture interior quality in a way that is completely independent of location.

We can also improve how the model handles geography by using advanced spatial clustering. Instead of feeding the model raw coordinate scalars or simple distance lines, applying algorithms like K-Means or DBSCAN, or joining the data with actual GIS layers, would allow us to group listings by meaningful boundaries like zip codes, school districts, or census tracts. Lastly, to solve the severe data imbalances between states, future pipelines should deploy algorithmic balancing frameworks like SMOTER for regression or strategic down-sampling. This will ensure that the final models maintain uniform predictive accuracy across all fifty states rather than biasing toward over-represented areas.

REFERENCES

- [1] Anders Bondemark and Axel Merkel. Parking not included: The effect of paid residential parking on housing prices and its relationship with public transport proximity. *Journal of Urban Economics*, 2023. <https://www.sciencedirect.com/science/article/pii/S0166046223000121>.
- [2] California Housing Affordability Tracker (1st Quarter 2025) [EconTax Blog]. California housing affordability tracker (1st quarter 2025). <https://lao.ca.gov/LA0EconTax/Article/Detail/793>, 2024. Accessed 2025-05-01.
- [3] Shixuan Cao, Weize Liao, and Jianrong Huang. Research on renting price prediction based on machine learning. *EAI Endorsed Transactions*, 2023. <https://eudl.eu/doi/10.4108/eai.8-12-2023.2344718>.
- [4] Shwetambari Chiwhane, Pooja Bagane, Akshansh Sourabh, Srushti Nagrale, Saakshi Jha, and Saumya Pandey. Estimarent: Data driven rental housing optimisation and market analysis for enhanced decision-making. *International Journal of Intelligent Systems and Applications in Engineering*, 12(2):20–28, 2023. <https://ijisae.org/index.php/IJISAE/article/view/4216>.
- [5] Sviatlana Engerstam. Long run apartment price dynamics in swedish and german cities. *Journal of European Real Estate Research*, 2021. <https://www.emerald.com/jerer/article/14/3/309/208354/Long-run-apartment-price-dynamics-in-Swedish-and?searchresult=1>.
- [6] Eun-Joo Han and Se-Hak Chun. A long short-term memory model using kernel density estimation for forecasting apartment prices in seoul city. *Expert Systems with Applications*, 2025. <https://www.sciencedirect.com/science/article/pii/S0957417425013703>.
- [7] LatLong.net. Reverse geocoding convert lat long to address. <https://www.latlong.net/Show-Latitude-Longitude.html>. Accessed 2026-06-29.
- [8] B. Manaswini, A. Sneha, M. Bharathi, and T. Aditya Sai Srinivas. Rentonomics: A machine learning approach to house rent prediction. *Zenodo Preprint*, 2024. <https://zenodo.org/records/10499800>.
- [9] David Mazáček and Jiří Panoš. Key determinants of sales price in the residential developments in prague. *Journal of European Real Estate Research*, 18(3):371–400, 2025. <https://www.sciencedirect.com/org/science/article/pii/S1753926925000099>.
- [10] Hajime Seya and Daiki Shiroi. A comparison of apartment rent price prediction using a large dataset: Kriging versus dnn. *arXiv preprint*, abs/1906.11099, 2019. Free PDF available at arXiv.
- [11] UCI Machine Learning Repository. Apartment for rent classified [dataset]. <https://doi.org/10.24432/C5X623>, 2019. Accessed 2025-05-01.

- [12] Kainuo Wang, Huiyi Zhao, and Jingzhong Li. Machine learning-based house rent prediction using stacking integration method. *American Journal of Mathematical and Computer Modeling*, 8(2):30–36, 2023. <https://www.sciencepg.com/article/10.11648/j.ajmse.20230802.12>.
- [13] Junhao Zhang. Predicting rental prices in india: A machine learning approach with hyperparameter optimization. *Advances in Economics, Management and Political Sciences*, 2024. <https://direct.ewa.pub/proceedings/aemps/article/view/18726>.
- [14] Zillow Rental Manager. California rental market: Average rental price in california & market trends. <http://www.zillow.com/rental-manager/market-trends/ca/?propertyTypes=apartment-condo>, 2025. Accessed 2025-05-20.