Comprehensive Data Mining Analysis of King County Housing Dataset

(STAT 6440 Final Project)

Innocent Abaa, Ebun Dosumu, Wanangwa Msiska, Daniel Sasu

Abstract

This study explores housing price prediction using the King County housing dataset, which contains 21,613 residential property sales recorded between May 2014 and May 2015 in King County, Washington. The primary goal is to identify key factors that influence house prices and to develop predictive models that can accurately estimate property values based on structural, locational, and historical features. Techniques used include multiple linear regression, LASSO, ridge regression, stepwise selection, regression trees, bagging, and random forests.

Clustering methods, including K-means, hierarchical, and model-based clustering, reveal hidden patterns in the data. Key predictors of price include living space and quality of material used in constructing the house (grade). Random forests provided the best predictive performance by capturing non-linear relationships.

Clustering uncovered distinct housing market segments, with K-means and hierarchical clustering identifying three main groups, while model-based clustering revealed nine submarkets. These findings suggest that house prices are influenced by size, age, and likely location.

Contents

1 Introduction

Housing prices play a pivotal role in shaping the economic stability and social well-being of individuals and communities. As one of the most significant financial assets for households, the value of a home is determined by a complex interplay of factors, including location, property characteristics, economic conditions, interest rates, and population growth. Accurate assessment and prediction of housing prices are essential for various stakeholders, including prospective buyers, real estate investors, policymakers, and urban planners. In recent years, advances in data mining and machine learning have significantly enhanced our ability to analyze large housing datasets and build reliable valuation models [7].

In this study, we expand on the idea of house price prediction by using the King County housing dataset as a case study to explore these relationships in greater detail. The dataset covers residential property transactions in King County, Washington, from May 2014 to May 2015. King County—home to major cities such as Seattle and Bellevue—is the most populous county in Washington State, with an estimated population of 2,052,800 in 2015 [6]. A higher population density typically corresponds with greater housing demand and more active real estate transactions, which in turn improves the quality and reliability of the available data.

The dataset, obtained from Kaggle and originally compiled by the King County Assessor's Office, contains $21,613$ house sales recorded in King County, with 21 various attributes describing the properties and their transactions [2].

Response Variable:

- **Price:** The sale price of a house (continuous variable).

Regressors (Predictor Variables):

- **Location-Based Features:** Latitude, longitude and zip code.
- **Structural Features:** Number of bedrooms and bathrooms, square footage of living space, square footage of the lot, number of floors, waterfront presence, view rating, condition rating, and grade.
- **Temporal Features:** Year the house was built, year of renovation.

- **Neighborhood Attributes:** Average size of interior housing living space for the closest 15 houses, Average size of land lots for the closest 15 houses, in square feet.

From a broader perspective, the application of data mining techniques in real estate analytics directly supports the goals of the United Nations Sustainable Development Goal (SDG) 11: Sustainable Cities and Communities, which emphasizes the importance of inclusive, safe, and affordable housing. By leveraging predictive modeling, this research contributes to data-driven decision-making, equitable housing policy, and efficient resource allocation. Additionally, it aligns with SDG 9: Industry, Innovation, and Infrastructure, as insights from predictive models can support infrastructure planning and strategic real estate development. Furthermore, it advances SDG 10: Reduced Inequalities by enabling a deeper understanding of real estate trends, which can inform policies aimed at promoting equitable access to housing across different regions and socio-economic groups.

Previous studies have successfully applied statistical and machine learning approaches—including multiple linear regression, decision trees, random forests, and gradient boosting machines—to predict housing prices and capture complex, nonlinear relationships between variables [1] [3]. In this study, we extend this body of work by exploring a wider range of modeling techniques. These include multiple linear regression, stepwise selection, LASSO regression, ridge regression, and ensemble methods, as well as unsupervised learning approaches such as K-means clustering, hierarchical and Model-based clustering.

The remainder of this report is organized as follows:

- **Section 2** describes the data preprocessing steps, identification of anomalies, results of exploratory data analysis and visualizations.
- **Section 3** discusses model development and performance evaluation metrics.
- **Section 4** discusses the results of the Analysis.
- **Section 5** summarizes the key findings, limitations and recommendations for future works.
- **Section 6** contains supplementary materials in Appendix.

2 Exploratory Data Analysis (EDA)

2.1 Data Cleaning

In this section, we cleaned the data from anomalies and explored some summary statistics (see Appendix for R outputs and more comprehensive details).

- The "Waterfront" variable was originally coded as integer, but we recoded it as a categorical variable ("0"-No waterfront, "1"-Waterfront).
- We created a new variable - **"yr_renovated"**. This represent the difference between the renovation year and the sale year to better reflect its impact on property value.
- We noted a few anomalies in the dataset:
  - we ensured the date variable is properly formatted.
  - There were 21 recorded transactions where either the year the house was built/renovated is greater than the year the sale occurred, meaning the house was sold before it was built or it was renovated after it was sold. Based on this, we excluded those transactions from the dataset.
  - The size of living area ("sqft_living") is the sum of living area above ("sqft_above") and size of living area in the basement ("sqft_basement"), so we would exclude the size of living area in the basement from the data to avoid perfect multicollinearity in our regression model.
  - The spread from the 3rd quartile for Price ($\$645,000$) to the maximum price ($\$7,700,000$) was so wide.
  - The spread from the 3rd quartile for bathrooms (4) to the maximum number of bathrooms (33) was so wide and 33 may be an error of commission.
  - Same suspicion with the Size of living area and Size of the lot variables too.

- So, we extracted all transactions with Bedrooms > 10, sqft_living > 10,000, sqft_lot > 1,000,000 to check for anomalies.
- Based on the result, we concluded that only the transaction with Bedroom = 33 was an anomaly because the house had only one floor, 1.75 bathroom and a small size of living area too. Therefore the observation was removed. We then proceed to obtain the final summary of the cleaned data.

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
price	75000	321500	450000	540098	645000	7700000
bedrooms	0.000	3.000	3.000	3.369	4.000	11.000
bathrooms	0.000	1.750	2.250	2.114	2.500	8.000
sqft_living	290	1428	1910	2080	2550	13540
sqft_lot	520	5040	7620	15114	10696	1651359
floors	1.000	1.000	1.500	1.494	2.000	3.500
waterfront	No :21431	Yes: 163				
view	0.0000	0.0000	0.0000	0.2339	0.0000	4.0000
condition	1.00	3.00	3.00	3.41	4.00	5.00
grade	1.000	7.000	7.000	7.656	8.000	13.000
sqft_above	290	1190	1560	1788	2210	9410
yr_built	1900	1951	1975	1971	1997	2015
yr_renovated	0.00	0.00	0.00	83.92	0.00	2015.00
sqft_living15	399	1490	1840	1987	2360	6210
sqft_lot15	651	5100	7620	12773	10084	871200
year_sold	2014	2014	2014	2014	2015	2015
house_age	0.00	18.00	40.00	43.33	63.00	115.00
age_renovated	0.00	16.00	37.00	40.97	60.00	115.00

Table 1: Summary of Cleaned data

2.2 Data Visualizations

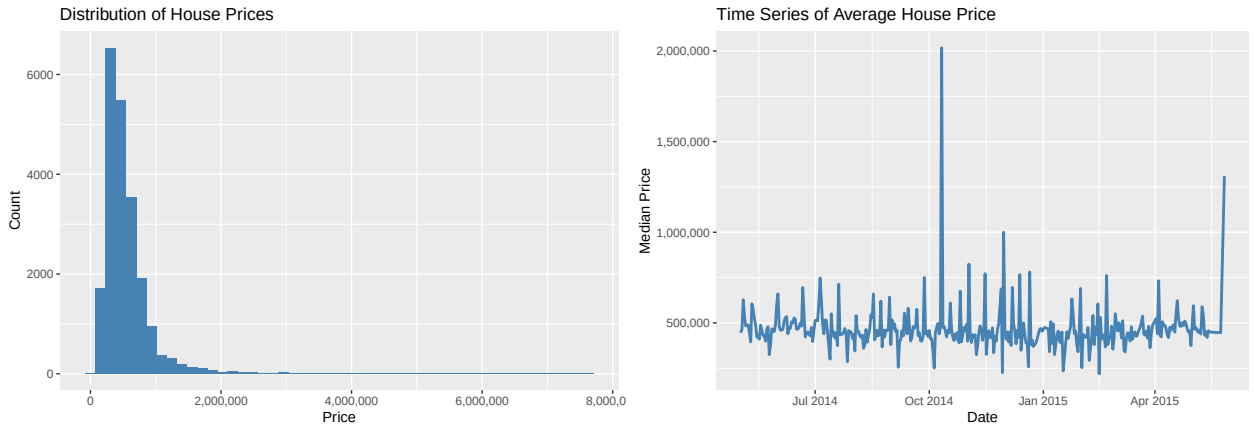
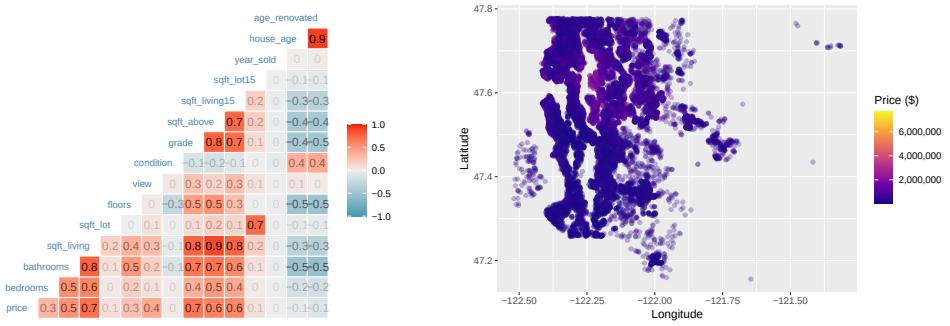


Figure 1: Histogram and Timesries Plot of House Prices

The histogram (Left) showed that the distribution of the house prices was right-skewed, indicating that there were a few houses that were very expensive compared to the majority of houses. Majority of the houses were priced between $\$300,000$ and $\$650,000$, with a few houses priced over $\$2,000,000$.

The timeseries plot shows that, on average, while the house prices appear to be stationary over time, there was a sudden jump in average house price between October and November 2014 before price returned to stationary and then jumped again in May 2015.



(a) Correlation plot for Numeric Variables

(b) Geographic Distribution of House Prices in King County

Figure 2: Housing Characteristics in King County

sqft_living	grade	sqft_above	sqft_living15	bathrooms	view	bedrooms	floors	sqft_lot	sqft_lot15
0.7024	0.6677	0.6058	0.5854	0.5254	0.3974	0.3156	0.2573	0.0896	0.0824

The square footage of living area and the house grade were the most correlated with house price. The higher the size of living area, on average, the higher the price.

Houses with higher prices were concentrated around Seattle and waterfront areas (latitudes near 47.6–47.7 and longitudes around -122.3 to -122.2). Lower-priced homes were more spread out in the south and east regions. This spatial trend aligns with urbanization and proximity to major amenities and coastlines.

(a) Distribution of Houses by Year Built

(b) Relationship Between Price and Living Area

(c) Variation in Price by Number of Bedrooms

(d) Effect of Waterfront Presence on House Prices

Figure 3: Visualizing Housing Trends — Distribution and Relationships Among Key Features

The plot in Figure 3a shows that most houses in the dataset were between 1 and 70 years old. Figure 3c illustrates a general trend of increasing median house prices with the number of bedrooms, suggesting that larger homes tend to be priced higher. However, there are notable exceptions where houses with fewer bedrooms still command high prices. Finally, Figure 3d reveals that properties with waterfront views have significantly higher median prices compared to those without, highlighting the premium associated with such features.

3 Methodology

This project employed a wide range of supervised and unsupervised learning techniques to model house prices in King County, Washington. Our methodological framework is structured in three main parts: **regression models**, **tree-based and ensemble methods**, and **clustering techniques**. Each method offers unique strengths for prediction, interpretation, or segmentation, and allows us to understand both the determinants and latent structure of housing prices.

3.1 Regression Models

3.1.1 Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is a fundamental supervised learning technique used to model the linear relationship between a dependent variable, in this case, house price, and multiple independent variables that represent various housing attributes. The primary goal of MLR is to understand how different features of a house contribute to variations in its price, and to develop a predictive model that can estimate prices for new observations based on these features.

In the context of MLR, the dependent variable Y is modeled as a linear combination of several predictor variables $X_1, X_2, \ldots, X_p$, along with an error term ε to account for random variability not explained by the predictors. The mathematical form of the model is given by:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon \tag{1}$$

Here, Y represents the outcome (house price), $X_1, X_2, \ldots, X_p$ are the predictors or independent variables, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_p$ are the coefficients that measure the effect of each predictor on the outcome, and ε denotes the random error term.

To apply MLR to the King County housing dataset, we begin by selecting relevant predictors such as square footage, number of bedrooms, and location-related variables. We then fit a linear regression model, interpret the estimated coefficients to assess the influence of each variable on house price, and examine diagnostic plots and statistical tests to evaluate the model's validity. A well-specified model can provide both insights into the factors driving house prices and accurate predictions for future listings.

3.1.2 Stepwise Regression

To improve the predictive accuracy and interpretability of the multiple linear regression model, we now perform **stepwise variable selection**. This process helps identify the most significant predictors by iteratively adding or removing variables based on criteria such as the Akaike Information Criterion (AIC).

We combine this with **k-fold cross-validation** to validate model performance and ensure that selected variables generalize well to unseen data. Cross-validation mitigates overfitting and provides a more robust estimate of model performance.

3.1.3 LASSO and Ridge Regression

LASSO regression enhances linear regression by introducing an L_1 penalty term to the loss function. This regularization technique encourages **sparse models** by shrinking some coefficients exactly to zero, effectively

performing variable selection and simplifying the model. By penalizing the absolute values of the coefficients, LASSO helps prevent overfitting while retaining only the most relevant predictors.

To fit a LASSO model, we use the `glmnet` package in R, which implements efficient algorithms for regularized regression.

In comparison, **Ridge Regression** minimizes the following objective function:

$$\hat{\beta}_{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{3}$$

Here, λ is a tuning parameter that controls the strength of the regularization. The second term is the L_2 penalty, which shrinks coefficients but does not set them to exactly zero. This helps in reducing model complexity and multicollinearity, but retains all predictors.

In contrast, **LASSO Regression** minimizes a similar objective function but with an L_1 penalty:

$$\hat{\beta}_{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

The key difference is that the L_1 penalty in LASSO tends to shrink some coefficients to exactly zero, thereby performing variable selection. The parameter λ again controls the trade-off between fitting the data and regularizing the model.

In the next step, we will fit a LASSO regression model using the `glmnet` package and assess its predictive performance through cross-validation.

3.1.4 K-Nearest Neighbors (KNN)

This is nonparametric method that predicts house prices based on the average prices of the k most similar properties. This approach is particularly useful for capturing local trends in real estate markets.

3.2 Tree-Based and Ensemble Models

3.2.1 Regression Tree

Regression Trees are recursive partitioning models that split the predictor space into non-overlapping regions. At each split, the algorithm chooses a feature and a threshold that minimizes the variance within the resulting partitions.

These models are interpretable and handle non-linear relationships and interactions well. However, single trees can be unstable and prone to overfitting, which motivates the use of ensemble techniques.

3.2.2 Bagging

Bagging, or **Bootstrap Aggregating**, is an ensemble learning method designed to improve the stability and accuracy of machine learning algorithms by reducing variance. It works by generating multiple versions of a training dataset using bootstrap sampling (sampling with replacement) and training a separate model on each version. The final prediction is obtained by aggregating the results, typically through averaging in regression tasks.

Bagging is especially effective with high-variance models such as decision trees, and it helps prevent overfitting by smoothing out the prediction surface. In this section, we apply bagging using the `caret`[4] package in R with a large number of trees to reduce the overall prediction error.

We will assess model performance using RMSE on test datasets, and compare the results to our previous linear and regularized regression model

3.2.3 Random Forest

Random Forest is an advanced ensemble learning method that builds on Bagging by incorporating **random feature selection** at each split in the decision trees. This added layer of randomness helps to **reduce correlation** between the individual trees, resulting in a more robust and accurate prediction model.

Unlike Bagging, where all predictors are considered for every split, Random Forest selects a random subset of predictors at each node, which typically improves performance by reducing overfitting and enhancing generalization. It is particularly effective in handling nonlinear relationships and interactions among predictors.

3.3 Clustering Methods

3.3.1 Hierarchical Clustering

Hierarchical clustering builds a hierarchy of clusters using an agglomerative (bottom-up) approach. It does not require a pre-specified number of clusters and produces a dendrogram that visualizes the nested grouping structure.

We apply hierarchical clustering on standardized variables selected by LASSO regression to explore natural groupings in the housing market and identify submarkets with similar pricing behaviors.

3.3.2 K-Means Clustering

K-Means clustering partitions the dataset into k clusters by minimizing within-cluster sum of squares. It requires choosing the number of clusters beforehand, which we determine using the elbow method or silhouette analysis.

This technique helps uncover latent price segments and patterns in the housing market, enabling potential geographic or design-based segmentation.

3.3.3 Model-Based Clustering (mclust)

Model-based clustering assumes that data points are generated from a finite mixture of Gaussian distributions, each representing a cluster. The mclust [5] package in R uses the Bayesian Information Criterion (BIC) to automatically select the optimal number of clusters and covariance structure.

4 Results and Discussion

First, we converted the house prices to thousands to allow for proper formatting, so all house prices are in $\$'000$.

4.1 Regression Results

4.1.1 Multiple Linear Regression

Table 3: Multiple Linear Regression results

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-977.2237	20.6207	-47.3905	0.0000
bedrooms	-41.8622	2.4944	-16.7825	0.0000
bathrooms	53.1489	4.1200	12.9001	0.0000
sqft_living	0.1662	0.0055	29.9487	0.0000
sqft_lot	0.0000	0.0001	0.4255	0.6705
floors	23.0475	4.4923	5.1304	0.0000
waterfront	593.1379	22.4896	26.3739	0.0000
view	43.5178	2.6890	16.1836	0.0000

	Estimate	Std. Error	t value	Pr(>|t|)
condition	17.2561	2.9532	5.8433	0.0000
grade	118.0672	2.6772	44.1012	0.0000
sqft_above	-0.0044	0.0054	-0.8193	0.4127
sqft_living15	0.0231	0.0043	5.4239	0.0000
sqft_lot15	-0.0006	0.0001	-6.1024	0.0000
house_age	3.9285	0.1461	26.8834	0.0000
age_renovated	-0.3113	0.1534	-2.0294	0.0424

The summary above shows that the lot size and size of living area above were not statistically significant at 5% level. This is possibly due to the fact that lot size and size of living area are highly correlated. Also, size of living area above and size of living area are highly correlated. We then proceed to check the presence of multicollinearity between the variables (only top 5 VIFs are shown):

sqft_living	bathrooms	sqft_lot	floors	bedrooms
8.5233	3.3487	1.9943	1.9506	1.7021

Based on the above, we observe the presence of multicollinearity and so, based on the correlation plot, we proceed to remove the size of living area above, age since the house was renovated and square footage of the lot from the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-976.0114	20.5338	-47.5319	0
bedrooms	-42.0387	2.4953	-16.8472	0
bathrooms	53.5204	4.0139	13.3339	0
sqft_living	0.1721	0.0039	44.0874	0
floors	20.2405	4.0571	4.9889	0
waterfront	590.2902	22.4732	26.2664	0
view	45.5773	2.6276	17.3458	0
condition	15.8548	2.8954	5.4758	0
grade	122.2262	2.5439	48.0473	0
sqft_lot15	-0.0005	0.0001	-7.6750	0
house_age	3.6508	0.0793	46.0351	0

sqft_living	bathrooms	bedrooms	floors	waterfront
4.21	3.17	1.7	1.59	1.19

We note that the values of the coefficient of determination remain essentially unchanged in the updated model. Specifically, the Multiple R-squared is $R^2 = 0.6559$ and the Adjusted R-squared is Adjusted $R^2 = 0.6557$, which are nearly identical to the values from the initial model: $R^2 = 0.6567$ and Adjusted $R^2 = 0.6563$. This suggests that the modifications to the model had minimal impact on its overall explanatory power.

Also, multicollinearity has been eliminated. Final model is:

$$\begin{aligned}\widehat{\text{price}}('000) = &- 976 - 42.04(\text{bedrooms}) + 53.52(\text{bathrooms}) + 0.17(\text{sqft\_living}) \\ &+ 20.24(\text{floors}) + 590.3(\text{waterfront}) + 45.58(\text{view}) + 15.85(\text{condition}) \\ &+ 122.2(\text{grade}) - 0.0005(\text{sqft\_lot15}) + 3.65(\text{house\_age})\end{aligned}$$

We assess the model's performance using the Root Mean Squared Error (RMSE) on both the training and test datasets. Lower RMSE values indicate better predictive accuracy.

The training and test RMSE values are approximately $213,417 and $221,680.80, respectively. These results indicate that, on average, the model's predicted house prices could deviate from the actual prices by as much as $221,680.80. Such a large error margin may lead to significant underestimation or overestimation, potentially resulting in predicted prices that are unrealistically low, or even negative in extreme cases.

Furthermore, the final coefficient of determination, $R^2 = 0.6559$, suggests that the model explains about 65.6% of the variance in house prices. While this may indicate a moderately good fit, it also highlights that a substantial portion of the variability remains unexplained. This relatively low R^2 value implies that the relationship between house price and the predictor variables may not be strictly linear. As a result, it may be worthwhile to explore non-linear modeling techniques or more flexible machine learning methods to improve predictive accuracy.

4.1.2 Stepwise Variable Selection

Below, we use stepwise selection based on AIC using the `stepAIC()` function from the **MASS** package, along with 10-fold cross-validation using the **caret** package.

The final multiple linear regression model selected via stepwise AIC included the following predictors: `bedrooms`, `bathrooms`, `sqft_living`, `floors`, `waterfront`, `view`, `condition`, `grade`, `sqft_living15`, `sqft_lot15`, `house_age`, and `age_renovated`.

The fitted equation is:

$$\begin{aligned}\widehat{\text{price}}('000) = &-975.66 - 41.88(\text{bedrooms}) + 53.69(\text{bathrooms}) + 0.163(\text{sqft\_living}) \\ &+ 21.47(\text{floors}) + 592.20(\text{waterfront}) + 43.94(\text{view}) + 17.40(\text{condition}) \\ &+ 117.85(\text{grade}) + 0.0223(\text{sqft\_living15}) - 0.0006(\text{sqft\_lot15}) \\ &+ 3.93(\text{house\_age}) - 0.31(\text{age\_renovated})\end{aligned}$$

All selected variables are statistically significant and contribute to explaining variability in house prices. Notably, `waterfront`, `grade`, and `sqft_living` have strong positive effects, while `bedrooms` and `sqft_lot15` have negative coefficients, potentially due to multicollinearity or interactions with other variables. Interestingly, `age_renovated` enters the model with a negative coefficient, suggesting that, all else equal, more recently renovated homes (lower values for age_renovated) tend to have higher prices.

This model reflects a more refined and potentially more predictive structure than the initial full model, as it retains only the most informative variables while eliminating those that contributed little to model performance.

The Root Mean Squared Error (RMSE) of the final model on the test dataset is approximately $221,470.80. This value is consistent with the previous model's test RMSE of $221,680.80, indicating that stepwise selection did not significantly reduce prediction error. This reinforces the need to explore more flexible or non-linear modeling approaches if the goal is to further reduce prediction error.

4.1.3 LASSO and Ridge Regression

In this next step, we will fit LASSO and Ridge regression models using the `glmnet` package and assess their predictive performance through cross-validation.

Table 7: Best Parameter for Lasso Regression

	alpha	lambda
8	1	0.8

The test RMSE of $221,456.60 from the LASSO regression did not show a significant improvement over the test RMSE of $221,680.70 from the ordinary least squares (OLS) regression. This suggests that, on average, predicted house prices may deviate by approximately $221,456 from their actual values. Such a large error margin may lead to substantial underestimation or overestimation of property values, potentially resulting in predictions that are unrealistically low—possibly even zero or negative in extreme cases.

Just like in the OLS regression model, the final LASSO model also excludes the variable `sqft_lot` and `age_renovated`, as their coefficients were shrunk to exactly zero. In addition, `sqft_above` was also eliminated by LASSO due to its limited contribution to model performance. This indicates that, in the presence of other predictive variables, both `sqft_lot` and `sqft_above` do not significantly improve the model's ability to explain variation in house prices and can be safely omitted to simplify the model.

The fitted LASSO equation is:

$$\begin{aligned}\widehat{\text{price}}('000) = &-968.94 - 39.46(\text{bedrooms}) + 52.86(\text{bathrooms}) + 0.162(\text{sqft\_living}) \\ &+ 20.03(\text{floors}) + 587.17(\text{waterfront}) + 44.14(\text{view}) + 15.14(\text{condition}) \\ &+ 118.15(\text{grade}) + 0.021(\text{sqft\_living15}) - 0.0005(\text{sqft\_lot15}) + 3.62(\text{house\_age})\end{aligned}$$

Table 8: Best Parameter for Ridge Regression

	alpha	lambda
100	0	0.1

The Ridge regression model produced a test RMSE of $222,700.40, which is slightly higher than the test RMSEs from both the OLS regression ($221,680.70) and the LASSO regression ($221,456.60). Like LASSO, Ridge adds a regularization penalty, but it uses an L_2 norm, which shrinks all coefficients toward zero without eliminating any of them entirely. As a result, all predictors remain in the model, albeit with smaller magnitudes.

The fitted Ridge regression equation is:

$$\begin{aligned}\widehat{\text{price}}('000) = &-907.28 - 30.98(\text{bedrooms}) + 59.96(\text{bathrooms}) + 0.13(\text{sqft\_living}) \\ &+ 15.42(\text{floors}) + 562.40(\text{waterfront}) + 50.55(\text{view}) + 18.27(\text{condition}) \\ &+ 104.77(\text{grade}) + 0.029(\text{sqft\_above}) + 0.036(\text{sqft\_living15}) \\ &- 0.0005(\text{sqft\_lot15}) + 2.94(\text{house\_age}) + 0.47(\text{age\_renovated})\end{aligned}$$

Unlike LASSO, Ridge regression retains all features in the model, including those that were dropped in the LASSO solution such as `sqft_lot` and `sqft_above`. Although this may be beneficial in settings where all variables are expected to carry some predictive value, the slightly higher RMSE and retained complexity suggest that Ridge may not be the most effective regularization technique in this case. Further exploration with non-linear models may help uncover deeper patterns in the data.

4.1.4 Lift chart

To further evaluate model performance, we compared the OLS, LASSO, and Ridge regression models using a lift chart.

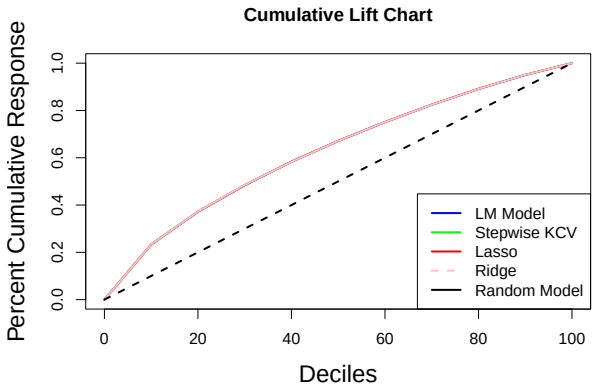


Figure 4: Lift Chart for Linear Regression Models

In our analysis, the three models—OLS, LASSO, and Ridge—performed **identically** in terms of lift. This is reflected in the lift chart, where all three model curves are **perfectly overlaid** behind the (yellow) line, indicating no substantial difference in their ability to rank-order predicted house prices.

This reinforces our earlier observations based on RMSE: while each model has subtle structural differences, their predictive power and ranking ability on this dataset are essentially equivalent.

4.2 K-Nearest Neighbor (KNN)

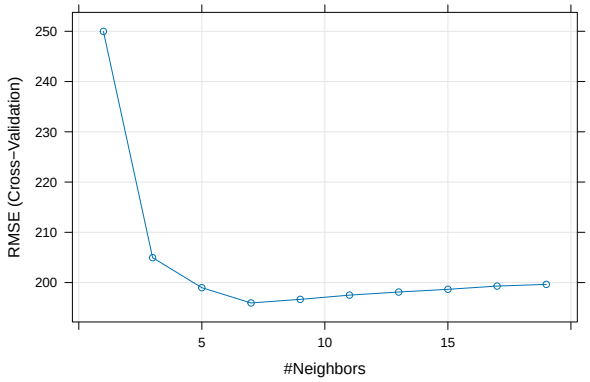


Figure 5: Cross-validation for KNN

The final model used was a 7-nearest neighbor regression model.

K-NN performed really poorly in terms of test RMSE ($\$659,129.90$), compared to the regression based models where the test RMSES were around ($\sim \$220,000$). This is due to the fact that K-NN uses distance metrics (usually Euclidean) to find the "nearest" neighbors. Due to the high-dimensionality of our data (with many housing features), distances become less meaningful due to the curse of dimensionality.

4.3 Tree-Based Models

4.3.1 Regression Tree

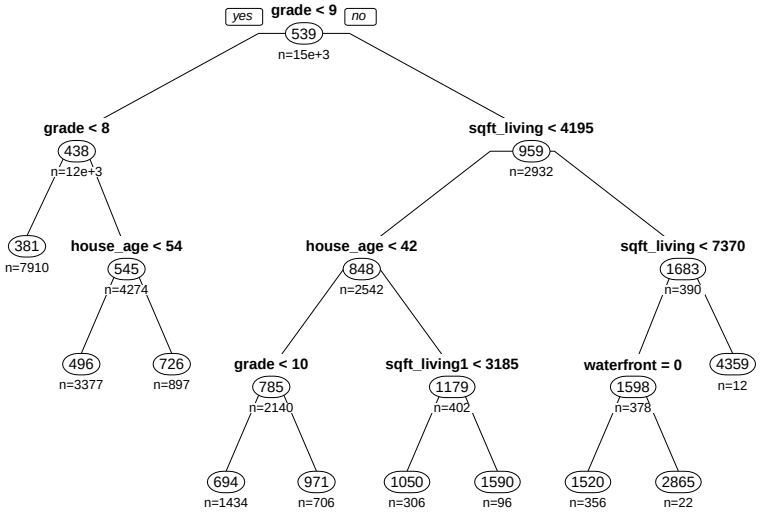
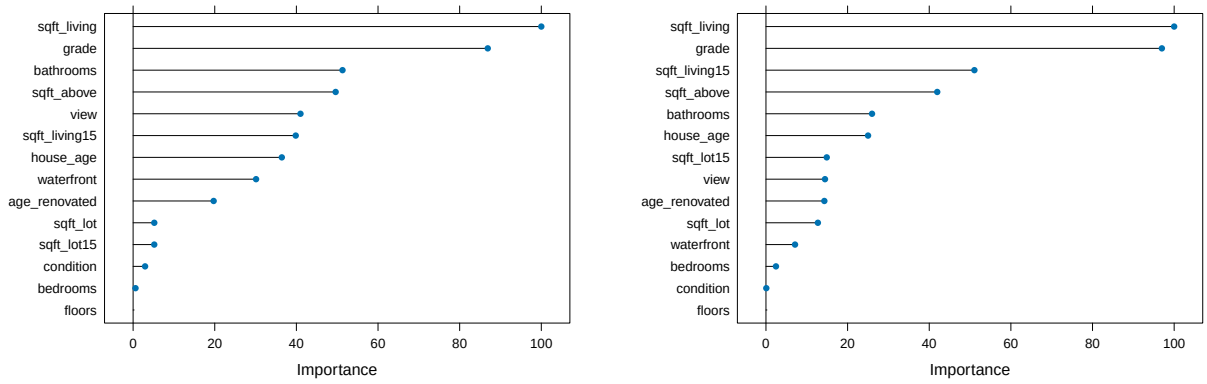


Figure 6: Regression Tree for house prices

Regression tree performed similarly (test RMSE = $\$248,301.80$), to OLS, Lasso and Ridge regression.

4.3.2 Bagging (Bootstrap Aggregating) and Random Forest



(a) Bagging

(b) Random Forest

Figure 7: Variable Importance Plots

The Bagging model not only improved prediction accuracy but also provided insight into the most influential predictors of house prices, as measured by their relative importance in reducing prediction error across the ensemble of trees.

The most important variable was `sqft_living`, which was standardized to a relative importance score of **100**. This indicates that the size of the living area is the most consistent and powerful predictor of house

price across all bootstrap samples. Following that, `grade` (86.88) and `bathrooms` (51.30) were also highly influential, suggesting that both the quality rating of the home and the number of bathrooms substantially affect price.

Other notable predictors include `sqft_above` (49.62), `view` (41.00), and `sqft_living15` (39.84), all of which capture aspects of house size, aesthetics, or surrounding features. Interestingly, `house_age` (36.43) and `waterfront` presence (30.12) also contributed significantly, further emphasizing the impact of both structural age and location on valuation.

Less influential variables included `age_renovated` (19.73) and `sqft_lot` (5.19), the latter of which consistently ranked low across previous models as well. These results highlight how Bagging helps identify the core set of variables driving house prices while mitigating overfitting by averaging over many decision trees.

The test RMSE for the Bagging model was approximately $227,448.70, which is slightly **higher** than the test RMSEs from the linear models. Despite its strength in reducing variance and capturing complex patterns, Bagging did not outperform the simpler linear models in this case. This could be due to the fact that the relationship between the predictors and the house price is relatively linear, and thus, linear models may suffice.

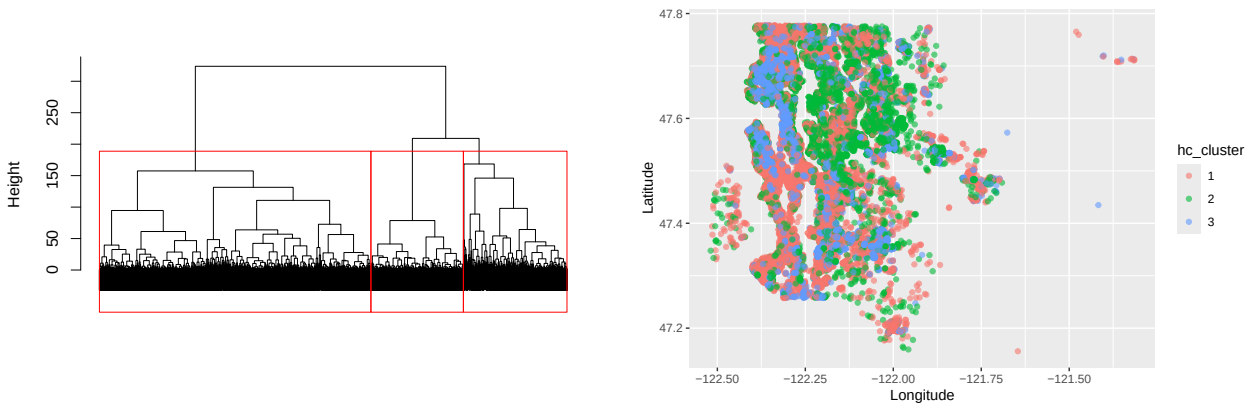
Nonetheless, Bagging provided valuable insights into variable importance, confirming that `sqft_living`, `grade`, and `number of bathrooms` are among the most critical predictors of house price. These variables consistently appeared as top contributors across all models, reinforcing their strong relationship with the response variable.

Random forest test RMSE was approximately $185,096.60, which is much **lower** than the test RMSEs from the linear models and an improvement over Bagging. The model identified similar most important variables (`sqft_living` and `grade`) with `grade` now having a higher importance score of 100 compared to 86.9 from Bagging.

4.4 Clustering Results

4.4.1 Hierarchical Clustering

We would use the variables selected by the Lasso regression to create a dendogram for the clusters.In addition, we employed "Ward's linkage" because unlike single/complete linkage, which are highly sensitive to outliers, Ward's method is less sensitive to outliers and focuses on overall structure, not just nearest neighbors.



(a) Dendogram based on Ward Linkage

(b) Clustering of Houses based on Location

Figure 8: Hierarchical Clustering

Table 9: Cluster Profile based on Hierarchical clustering

hc_cluster	price_mean	sqft_living_mean	bedrooms_mean	bathrooms_mean	house_age_mean
1	429152.2	1687.625	3.152006	1.715946	58.31135
2	919098.0	3235.185	4.077341	2.882055	30.27237
3	441294.0	1935.741	3.214236	2.424257	13.98759

The dendogram suggests that 3 clusters (types) of houses exists in the King County region.

- **Cluster 1 (small older houses)**: These are the most modest in size and price. The average house price is approximately $\$429k$, with an average living space of 1,688 square feet. These homes typically have 3 bedrooms and 1.7 bathrooms, and are about 58 years old, suggesting they were built or last renovated several decades ago.
- **Cluster 2 (expensive large-size, mid-age house)**: This group includes the most expensive and spacious homes. The average price is roughly $\$919k$, and the average living area is around 3,235 square feet. These houses typically feature 4 bedrooms and nearly 3 bathrooms, and are about 30 years old, indicating they are relatively modern, possibly in affluent areas such as Seattle or near waterfronts.
- **Cluster 3 (moderate-size, fairly new house)**: Homes in this group strike a balance between size and affordability. They have an average price of $\$441k$ and a living area of 1,936 square feet. With around 3 bedrooms and 2.4 bathrooms, these are the newest on average, with a mean age of 14 years, making them attractive for buyers seeking more recent builds at a moderate price.

4.4.2 K-Means Clustering

We use the number of $k = 3$ from the heirarchical clustering results to run a KMeans algorithm:

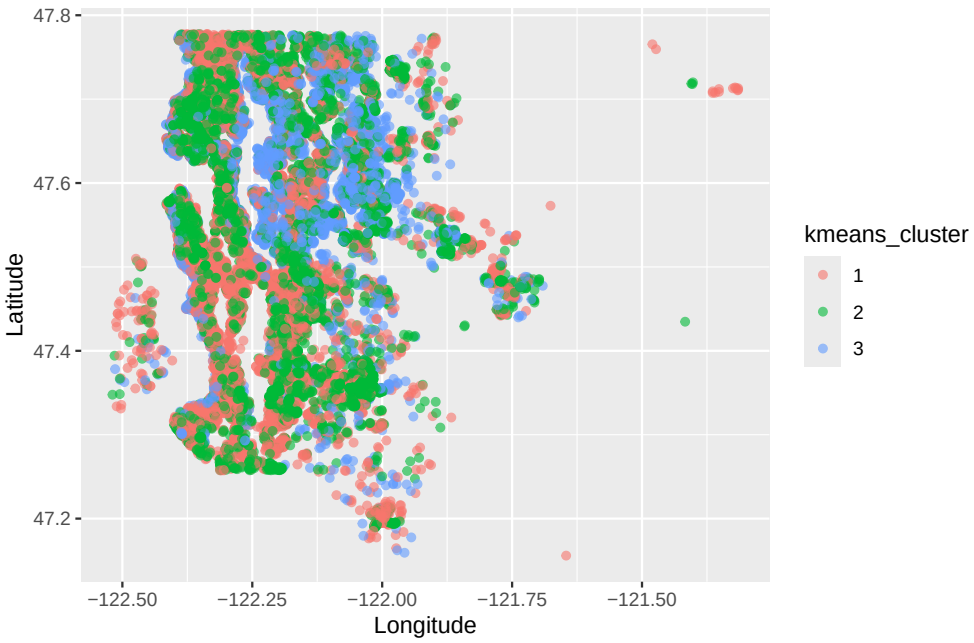


Figure 9: K-Means Clustering of Houses based on Location

Table 10: Cluster Profile based on K-Means clustering

kmeans_cluster	price_mean	sqft_living_mean	bedrooms_mean	bathrooms_mean	house_age_mean
1	405984.2	1551.917	3.041404	1.575584	62.12384
2	496935.4	2217.688	3.546626	2.499871	21.70647
3	1127771.5	3603.555	4.082400	3.045155	31.81015

- **Cluster 1 (small older houses)**: These homes are generally more affordable and compact. The average house price is approximately $\$406k$, with an average living area of 1,552 square feet. They typically offer 3 bedrooms and 1.6 bathrooms, and are the oldest group, averaging 62 years in age—indicating many of these homes may benefit from renovation or modernization.
- **Cluster 2 (moderate-size, newer house)**: This cluster represents homes with a good blend of size, modernity, and affordability. The average price is around $\$497k$, and the average size is 2,218 square feet. These homes generally have 3.5 bedrooms, 2.5 bathrooms, and are about 22 years old, suggesting they are relatively recent builds, possibly situated in suburban developments with good resale value.
- **Cluster 3 (expensive large-size, fairly new premium house)**: Homes in this cluster are spacious, high-end, and expensive, with an average price of approximately $\$1.13M$ — nearly triple that of Cluster 1. These homes average 3,604 square feet in size, come with 4 bedrooms and 3 bathrooms, and are about 32 years old. Their scale and price suggest they may be located in sought-after neighborhoods like Bellevue or waterfront areas, aligning with patterns often observed in premium market segments.

4.4.3 Model-Based Clustering

Model-based clustering assumes data is generated from a mixture of Gaussian distributions. It automatically selects the best number of clusters using BIC.

Table 11: Mclust number of observations per cluster

1	2	3	4	5	6	7	8	9
1626	1139	743	1725	3880	562	5610	3975	2335

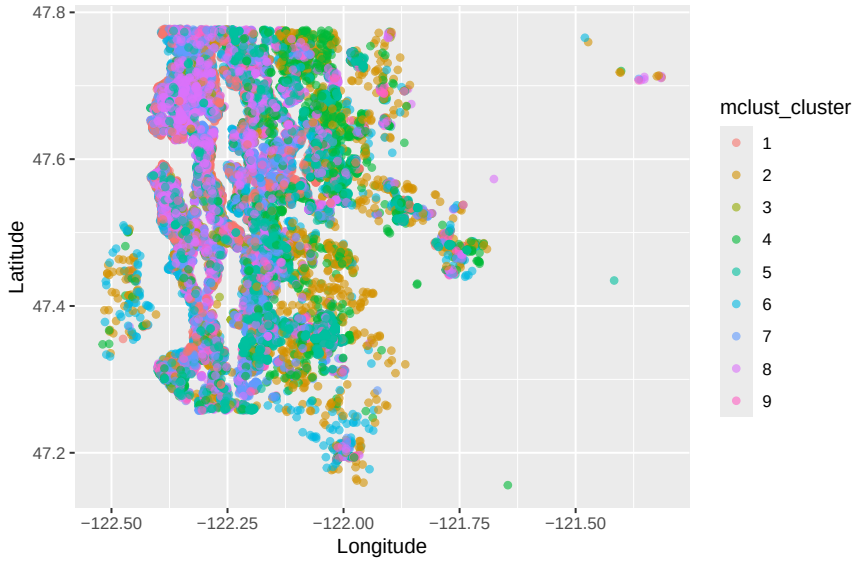


Figure 10: Model-based Clustering of Houses based on Location

Table 12: Cluster Profile based on Model-based clustering

mclust_cluster	price_mean	sqft_living_mean	bedrooms_mean	bathrooms_mean	house_age_mean
1	928155.0	2800.1212	3.642681	2.5272140	50.90037
2	480677.4	2253.1698	3.319929	2.0131814	47.11260
3	332618.2	988.7056	2.000000	0.9973118	74.09946
4	704434.8	3007.0420	3.797203	2.7808858	23.99476
5	479363.4	2279.1326	3.510241	2.4934076	13.60164
6	1091025.4	3093.6886	3.548043	2.6178826	43.47509
7	470572.8	1705.2580	3.258533	1.7372912	65.55592
8	554998.8	2046.7621	3.546864	2.2564576	37.79336
9	351788.9	1525.6533	3.000000	1.6311755	47.55920

The model-based clustering identified 9 distinct segment of homes in the King County region based on housing characteristics. These clusters capture meaningful variations in price, size, layout, and age of the properties. Below is a summary of the key characteristics of each group:

- **Cluster 1 (expensive, large mid-age homes)**: Homes in this group have an average price of $\$928k$ and 2,800 sqft of living space, with around 3.6 bedrooms, 2.5 bathrooms, and are typically 51 years old. Likely located in desirable neighborhoods with upgraded features.
- **Cluster 2 (moderate-price, mid-size homes)**: Averaging $\$481k$, these homes offer 2,253 sqft, 3.3 bedrooms, and 2 bathrooms, with an average age of 47 years. They strike a balance between size and affordability.
- **Cluster 3 (small, older homes)**: The most affordable group, priced at $\$333k$ on average. These homes are compact (989 sqft), with 2 bedrooms, 1 bathroom, and are the oldest, averaging 74 years in age.
- **Cluster 4 (large, newer homes)**: High-value properties averaging $\$704k$ and 3,007 sqft, with 3.8 bedrooms, 2.8 bathrooms, and a relatively young average age of 24 years. Likely to feature modern designs and amenities.
- **Cluster 5 (mid-size, newer homes)**: With an average price of $\$479k$, these homes have 2,279 sqft, 3.5 bedrooms, 2.5 bathrooms, and are relatively new at 14 years old, suggesting recent construction or renovation.
- **Cluster 6 (premium homes)**: The most expensive cluster with homes averaging $\$1.09M$, 3,094 sqft, 3.5 bedrooms, 2.6 bathrooms, and around 43 years old. Likely reflects upscale, well-maintained properties in prime areas.
- **Cluster 7 (affordable mid-size older homes)**: Priced around $\$471k$, these homes have 1,705 sqft, 3.3 bedrooms, 1.7 bathrooms, and are 66 years old on average — possibly older suburban developments.
- **Cluster 8 (mid-size, mid-age homes)**: Averaging $\$555k$, these homes offer 2,047 sqft, 3.5 bedrooms, and 2.3 bathrooms, with an average age of 38 years. Mid-range options with potential for value appreciation..
- **Cluster 9 (affordable, moderate-size, mid-age homes)**: With a price tag of $\$352k$, these houses have 1,526 sqft, 3 bedrooms, 1.6 bathrooms, and are 47 years old. These might represent entry-level homes in stable, older neighborhoods.

This clustering reveals distinct market niches—from compact vintage homes to spacious newer builds. Notably, price is not solely driven by size, but also by age and likely location, which clustering helps reveal.

For instance, Cluster 1 and Cluster 6 have similarly high price points (around $\$928k$ and $\$1.09M$, respectively), yet differ in both size and age — Cluster 6 homes are slightly larger (3,093 sqft vs. 2,800 sqft) but somewhat newer (43 vs. 51 years), potentially reflecting premium locations or renovations. On the other hand, Cluster 3 contains the smallest and oldest homes (988 sqft, 74 years old) with the lowest average price of $\$332k$, indicating a budget-friendly segment. Meanwhile, Cluster 5 stands out as the youngest cluster (only ~ 14 years old), yet maintains a moderate price point of $479k, appealing to value-conscious buyers seeking newer construction without the premium of upscale neighborhoods.

5 Discussion and Conclusion

This study explored various data mining techniques to model and understand house prices in King County, Washington. The analysis combined both supervised learning (regression models) and unsupervised learning (clustering techniques) to gain insights from the housing dataset.

Model	Test RMSE
Multiple Linear Regression	$221,680.80
Stepwise Variable Selection	$221,470.80
LASSO Regression	$221,456.60
Ridge Regression	$222,700.40
K-Nearest Neighbors (K-NN)	$659,129.90
Regression Tree	$248,301.80
Bagging	$227,448.70
Random Forest	$184,928.30

Table 13: Comparison of Test RMSE Across supervised learning Models

5.1 Key Findings

Among the models evaluated, Random Forest Regression emerged as the best-performing model in terms of predictive accuracy, achieving a test RMSE of approximately $\sim \$184,000$. This performance significantly outperformed simpler models like k-Nearest Neighbors (K-NN), which had a test RMSE exceeding $659,000, likely due to the sensitivity of KNN to high-dimensional spaces and lack of model interpretability.

The Bagging approach, as an ensemble method, also performed robustly, reducing variance and improving stability over single regression trees. Regression Trees alone provided interpretable decision rules, highlighting key split points (e.g., sqft_living and grade), but they lacked the predictive strength of ensemble methods.

On the unsupervised front, hierarchical clustering using Ward's linkage revealed meaningful groupings in the housing data. Ward's method was particularly effective due to its ability to create compact, spherical clusters by minimizing within-cluster variance. This allowed for a more balanced and interpretable cluster structure compared to other linkage methods such as single or complete linkage, which are more sensitive to outliers or chaining effects. The insights from Model-based clustering highlight how the method uncovers hidden patterns, offering a nuanced, data-driven view of housing submarkets across King County.

5.2 Limitations

Several limitations impacted the analysis:

- **Computational Cost:** The Random Forest model, while accurate, required significant computation time, especially with a large number of trees and features. Similarly, hierarchical clustering—particularly with Ward's method—required computing a full distance matrix, which can be computationally intensive with larger datasets like ours.
- **Data Constraints:** The dataset is cross-sectional and limited to homes sold in a specific period. Important temporal dynamics such as seasonality, market trends, or future changes in housing value could not be captured.
- **Model Assumptions:** Some models (e.g., linear regression) rely on assumptions such as linearity, homoscedasticity, and independence of errors, which may not fully hold in this context.
- **Feature Limitations:** Some potentially impactful variables like neighborhood amenities, school district quality were not available in the dataset.

5.3 Recommendations and Future Directions

Based on the findings, several avenues for future work are recommended:

- **Time Series Forecasting:** Incorporating temporal dynamics into the analysis through time series models would provide better understanding of market trends and future price movements.
- **Spatial Modeling:** Given the importance of location in real estate, spatial models could better capture locational effects.
- **Enhanced Feature Engineering:** Incorporating domain-specific features or using external datasets (e.g., school ratings, proximity to public transport) could improve model accuracy and interpretability.

6 Appendix/Supplementary Materials

- **Data Structure**

```
## [1] 21613    21
```

```
## 'data.frame':    21613 obs. of  21 variables:
##  $ id           : num  7129300520 6414100192 5631500400 2487200875 1954400510 ...
##  $ date         : chr  "20141013T000000" "20141209T000000" "20150225T000000" "20141209T000000" ...
##  $ price        : num  221900 538000 180000 604000 510000 ...
##  $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
##  $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
##  $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
##  $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
##  $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
##  $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ condition    : int  3 3 3 5 3 3 3 3 3 3 ...
##  $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
##  $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
##  $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
##  $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
##  $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
##  $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
##  $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
##  $ long         : num  -122 -122 -122 -122 -122 ...
##  $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
##  $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
```

- **Check for Missing data**

```
##            id          date         price      bedrooms     bathrooms
##             0             0             0             0             0
##   sqft_living      sqft_lot        floors    waterfront          view
##             0             0             0             0             0
##     condition         grade    sqft_above sqft_basement      yr_built
##             0             0             0             0             0
##  yr_renovated       zipcode           lat          long sqft_living15
##             0             0             0             0             0
##    sqft_lot15
##             0
```

- **Check for Anomalies**

	id	date	price	yr_built	yr_renovated
1764	1832100030	2014-06-25	597326	2015	0
2688	3076500830	2014-10-29	385195	2015	0

	id	date	price	yr_built	yr_renovated
7527	9520900210	2014-12-31	614285	2015	0
8040	1250200495	2014-06-24	455000	2015	0
14490	2770601530	2014-08-26	500000	2015	0
17099	9126100346	2014-06-17	350000	2015	0
19806	9126100765	2014-08-01	455000	2015	0
20771	9310300160	2014-08-28	357000	2015	0
20853	1257201420	2014-07-09	595000	2015	0
20964	6058600220	2014-07-31	230000	2015	0
21263	5694500840	2014-11-25	559000	2015	0
21373	6169901185	2014-05-20	490000	2015	0
2296	8712100320	2014-07-28	585000	1922	2015
7098	9141100005	2014-10-28	285000	1940	2015
11600	7284900030	2014-05-22	850000	1923	2015
14860	3585900665	2014-06-06	805000	1956	2015
15688	3585900190	2014-10-06	825000	1955	2015
18576	8935100100	2014-07-01	476000	1945	2015

We noticed that there were 21 observations where the year of sale was after the year the house were built.

- **Remove anomalies, Transaction ID, and sqft_basement because of perfect correlation with sqft_living**
- **Get summary of data and check for more anomalies**

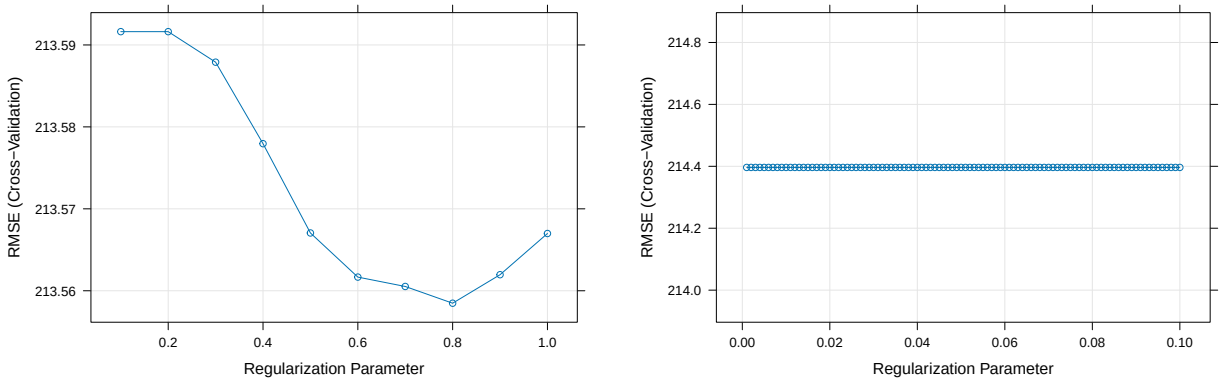
```
##       date                price             bedrooms        bathrooms
##  Min.   :2014-05-02   Min.   :  75000   Min.   : 0.000   Min.   :0.000
##  1st Qu.:2014-07-22   1st Qu.: 321500   1st Qu.: 3.000   1st Qu.:1.750
##  Median :2014-10-16   Median : 450000   Median : 3.000   Median :2.250
##  Mean   :2014-10-29   Mean   : 540102   Mean   : 3.371   Mean   :2.114
##  3rd Qu.:2015-02-17   3rd Qu.: 645000   3rd Qu.: 4.000   3rd Qu.:2.500
##  Max.   :2015-05-27   Max.   :7700000   Max.   :33.000   Max.   :8.000
##   sqft_living       sqft_lot           floors      waterfront       view
##  Min.   :  290   Min.   :    520   Min.   :1.000   No :21432   Min.   :0.0000
##  1st Qu.: 1428   1st Qu.:   5041   1st Qu.:1.000   Yes:  163   1st Qu.:0.0000
##  Median : 1910   Median :   7620   Median :1.500               Median :0.0000
##  Mean   : 2080   Mean   :  15113   Mean   :1.494               Mean   :0.2339
##  3rd Qu.: 2550   3rd Qu.:  10696   3rd Qu.:2.000               3rd Qu.:0.0000
##  Max.   :13540   Max.   :1651359   Max.   :3.500               Max.   :4.0000
##    condition        grade          sqft_above      yr_built
##  Min.   :1.00   Min.   : 1.000   Min.   : 290   Min.   :1900
##  1st Qu.:3.00   1st Qu.: 7.000   1st Qu.:1190   1st Qu.:1951
##  Median :3.00   Median : 7.000   Median :1560   Median :1975
##  Mean   :3.41   Mean   : 7.656   Mean   :1788   Mean   :1971
##  3rd Qu.:4.00   3rd Qu.: 8.000   3rd Qu.:2210   3rd Qu.:1997
##  Max.   :5.00   Max.   :13.000   Max.   :9410   Max.   :2015
##   yr_renovated        zipcode           lat             long
##  Min.   :   0.00   Min.   :98001   Min.   :47.16   Min.   :-122.5
##  1st Qu.:   0.00   1st Qu.:98033   1st Qu.:47.47   1st Qu.:-122.3
##  Median :   0.00   Median :98065   Median :47.57   Median :-122.2
##  Mean   :  83.91   Mean   :98078   Mean   :47.56   Mean   :-122.2
##  3rd Qu.:   0.00   3rd Qu.:98117   3rd Qu.:47.68   3rd Qu.:-122.1
##  Max.   :2015.00   Max.   :98199   Max.   :47.78   Max.   :-121.3
##  sqft_living15    sqft_lot15
```

```
##  Min.   : 399   Min.   :   651
##  1st Qu.:1490   1st Qu.:  5100
##  Median :1840   Median :  7620
##  Mean   :1987   Mean   : 12773
##  3rd Qu.:2360   3rd Qu.: 10084
##  Max.   :6210   Max.   :871200
```

	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors
1720	2015-03-27	700000	4	1.00	1300	1651359	1.0
7648	2014-05-21	542500	5	3.25	3010	1074218	1.5
7770	2015-01-19	855000	4	3.50	4030	1024068	2.0
17320	2015-05-04	190000	2	1.00	710	1164794	1.0
3915	2014-06-11	7062500	5	4.50	10040	37325	2.0
7253	2014-10-13	7700000	6	8.00	12050	27600	2.5
12778	2014-05-05	2280000	7	8.00	13540	307752	3.0
8758	2014-08-21	520000	11	3.00	3000	4960	2.0
15871	2014-06-25	640000	33	1.75	1620	6000	1.0
1165	2014-10-20	5110800	5	5.25	8010	45517	2.0
1316	2015-04-13	5300000	6	6.00	7390	24829	2.0
1449	2015-04-13	5350000	5	5.00	8000	23985	2.0
3915.1	2014-06-11	7062500	5	4.50	10040	37325	2.0
4412	2014-08-04	5570000	5	5.75	9200	35069	2.0
7253.1	2014-10-13	7700000	6	8.00	12050	27600	2.5
9255	2014-09-19	6885000	6	7.75	9890	31374	2.0

We concluded that only the transaction with Bedroom = 33 was an anomaly because the house had only one floor, 1.75 bathroom and a small size of living area too. Therefore the observation was removed and we now have our final data.

- **RMSE Cross-validation Plot for Lasso and Ridge Models**



(a) Lasso Model

(b) Ridge Model

Figure 11: Cross-validation Plot

References

[1] Evgeny A. Antipov and Elena B. Pokryshevskaya. "Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a CART-Based Approach for Model Diagnostics". In: *Expert Systems with Applications* 39.2 (2012), pp. 1772–1778. DOI: 10.1016/j.eswa.2011.08.077.

[2] GeoDa Center for Geospatial Analysis and Computation. *King County House Sales Data*. 2025. URL: https://geodacenter.github.io/data-and-lab/KingCounty-HouseSales2015/.

[3] Ahmed Khamis, Rasha Hussein, and Heba Mohamed. "A Comparative Study of Machine Learning Algorithms for Predicting House Prices". In: *International Journal of Advanced Computer Science and Applications* 11.1 (2020), pp. 390–397. DOI: 10.14569/IJACSA.2020.0110151.

[4] Kuhn and Max. "Building Predictive Models in R Using the caret Package". In: *Journal of Statistical Software* 28.5 (2008), pp. 1–26. DOI: 10.18637/jss.v028.i05. URL: https://www.jstatsoft.org/index.php/jss/article/view/v028i05.

[5] Luca Scrucca et al. *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. Chapman and Hall/CRC, 2023. ISBN: 978-1032234953. DOI: 10.1201/9781003277965. URL: https://mclust-org.github.io/book/.

[6] Washington State Office of Financial Management. *April 1 Population Estimates of Cities, Towns and Counties*. Retrieved from https://ofm.wa.gov/. 2015.

[7] Yu Zhang, Yu Zheng, and Bo Li. "Urban Computing: Concepts, Methodologies, and Applications". In: *ACM Transactions on Intelligent Systems and Technology* 9.4 (2018), pp. 1–25. DOI: 10.1145/3213344.