Understanding the Spatial Patterns and Modeling Drivers of Housing Prices in King County, Washington

Chenyi Weng

SSCI575 - Spatial Data Science

Professor Yi Qi

October 28, 2025

Part 1 – Exploratory Spatial Data Analysis (ESDA)

1(a) Spatial Autocorrelation of Housing Prices in King County

To examine whether housing prices in King County exhibit spatial dependence, both global and local spatial autocorrelation analyses were conducted using ArcGIS Pro. The dataset (kc_house_data) was first projected to the *NAD 1983 UTM Zone 10N* coordinate system to ensure that all distance-based calculations were made in meters. The variable price was used as the input field for all spatial statistical tools.

The Global Moran's I statistic was calculated using the *K Nearest Neighbors* conceptualization with eight neighbors. The resulting Moran's I value was 0.591, with a z-score of 183.77 and a p-value < 0.001, indicating a highly significant positive spatial autocorrelation (Figure 1). This means that similar house prices—either high or low—tend to cluster together rather than being randomly distributed. The pattern suggests that property values in King County are spatially structured, with local environments and neighborhood characteristics influencing market prices.

Global Moran's I Summary

Moran's Index	0.591132
Expected Index	-0.000046
Variance	0.000010
z-score	183.766717
p-value	0.000000

Succeeded at Sunday, October 26, 2025 4:01:52 PM (Elapsed Time: 6.76 seconds)

Figure 1. Global Moran's I Summary of Housing Prices in King County, WA.

To further identify where these spatial clusters occur, the Anselin Local Moran's I (Cluster and Outlier Analysis) was applied with the same parameters (K = 8 neighbors, 499 permutations). The results revealed clear geographic patterns of housing price clusters across King County (Figure 2). High–High clusters, shown in red, are primarily located around Seattle, Bellevue, and Kirkland, indicating concentrations of high-value homes in urban and waterfront areas with strong economic activity. Low–Low clusters, shown in blue, dominate the southern and eastern suburbs, representing neighborhoods with relatively lower property values. High–Low (yellow) and Low–High (green) outliers appear at the edges of these clusters, signifying isolated anomalies where expensive houses are surrounded by cheaper properties or vice versa.

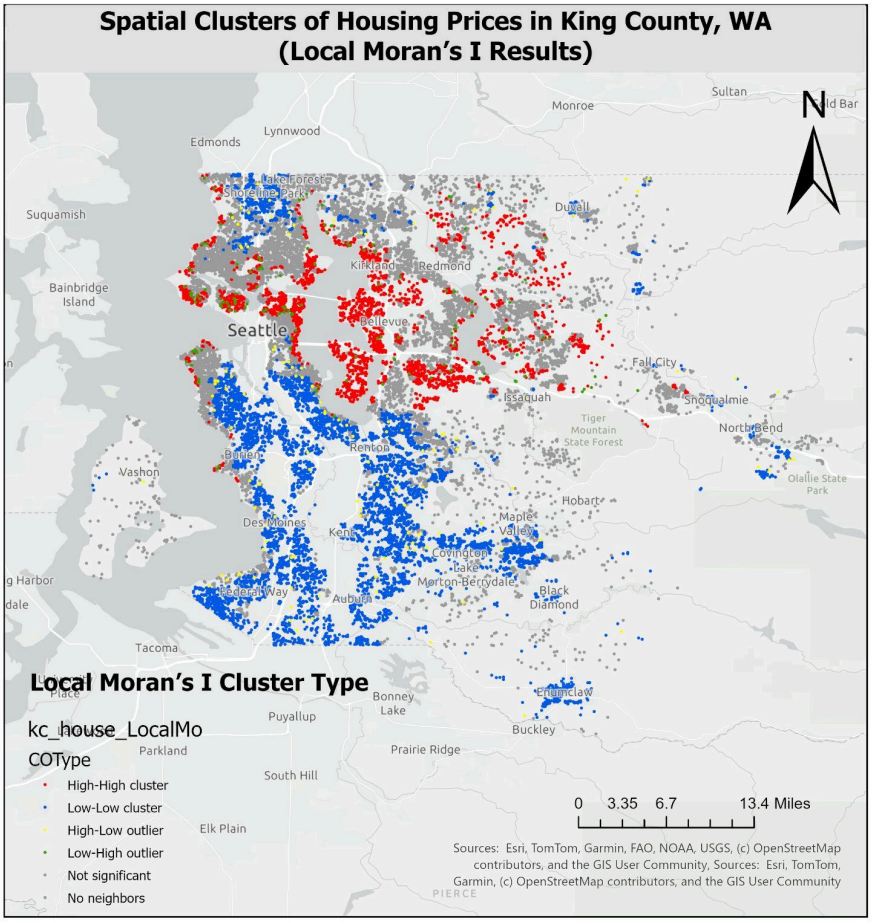


Figure 2. Spatial Clusters of Housing Prices in King County, WA (Local Moran's I Results).

Overall, the combination of global and local spatial analyses confirms that housing prices in King County are not randomly distributed but exhibit strong and statistically significant spatial clustering. These spatial patterns reflect the underlying socioeconomic and geographic heterogeneity of the region, such as accessibility to employment centers, proximity to Lake Washington, and differences in land use intensity between urban cores and peripheral areas.

1(b) Visualizing Relationships between Housing Price and Other Variables

To further understand the factors influencing housing prices in King County, several charts were created using the projected dataset *kc_house_data_UTM* to visualize relationships between house price and different property attributes. Scatter plots were used for continuous variables such as living area, number of bedrooms, number of bathrooms, and grade. Box plots were applied for categorical variables such as waterfront presence and view score.

The first scatter plot (Figure 3) shows the relationship between **living area (sqft_living)** and **house price**. A clear positive trend can be observed, with an R^2 value of 0.49, indicating that larger houses tend to have higher prices. The second scatter plot (Figure 4) explores the relationship between **bedrooms** and **price**, but the R^2 value of 0.10 suggests that the number of bedrooms alone is a weak predictor of price. A third scatter plot (Figure 5) plots **bathrooms** against **price**, with an R^2 of 0.28, showing a moderate positive relationship between the number of bathrooms and housing price. The fourth scatter plot (Figure 6) presents **grade** versus **price**, where the R^2 value of 0.45 indicates that homes with higher construction and design quality (grade) generally have higher prices.

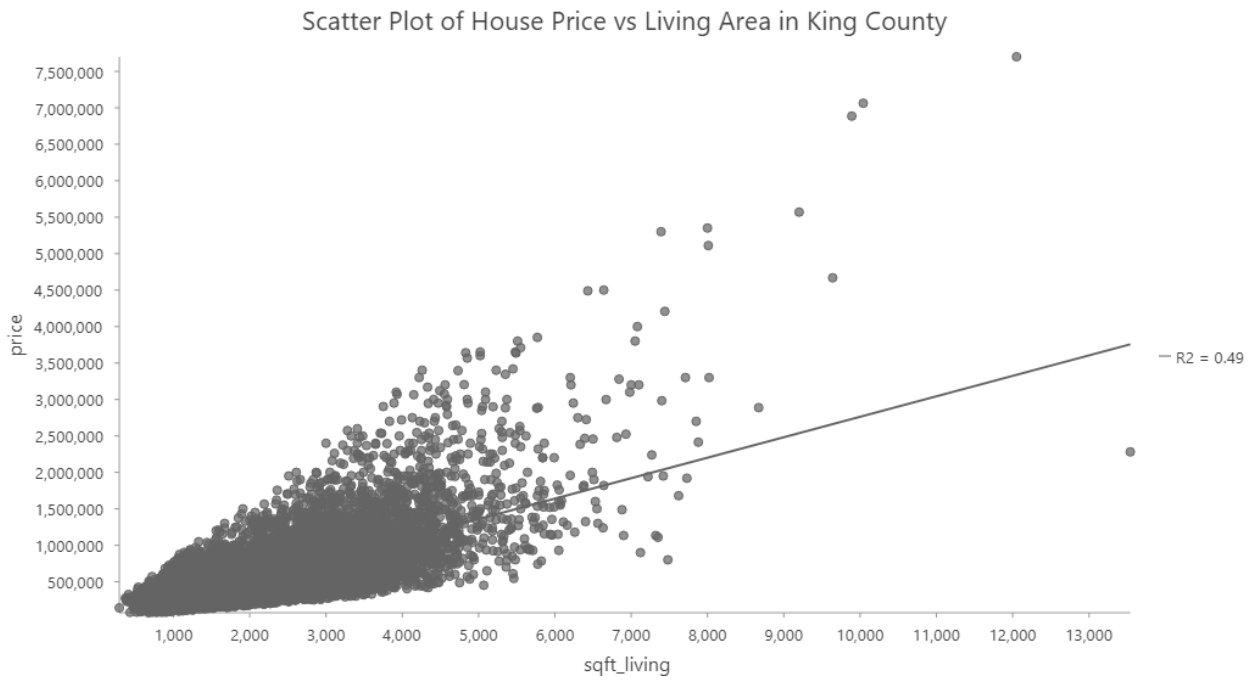


Figure 3: Scatter Plot of House Price vs Living Area in King County

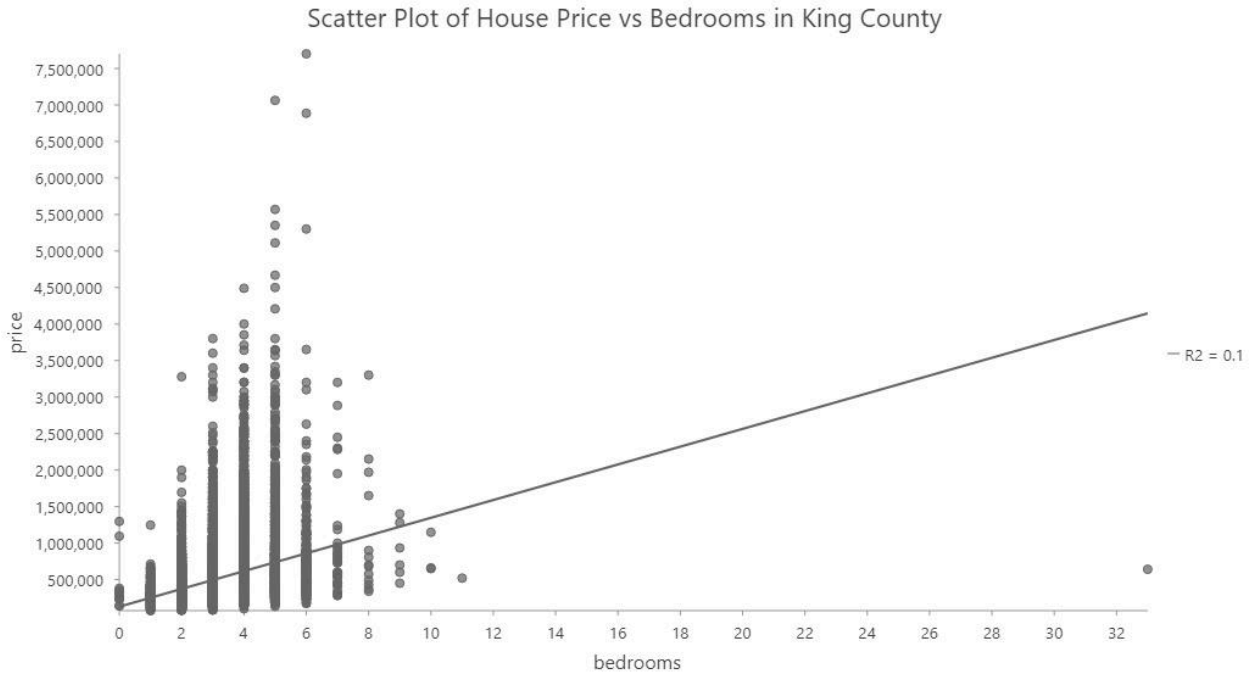


Figure 4: Scatter Plot of House Price vs Bedrooms in King County

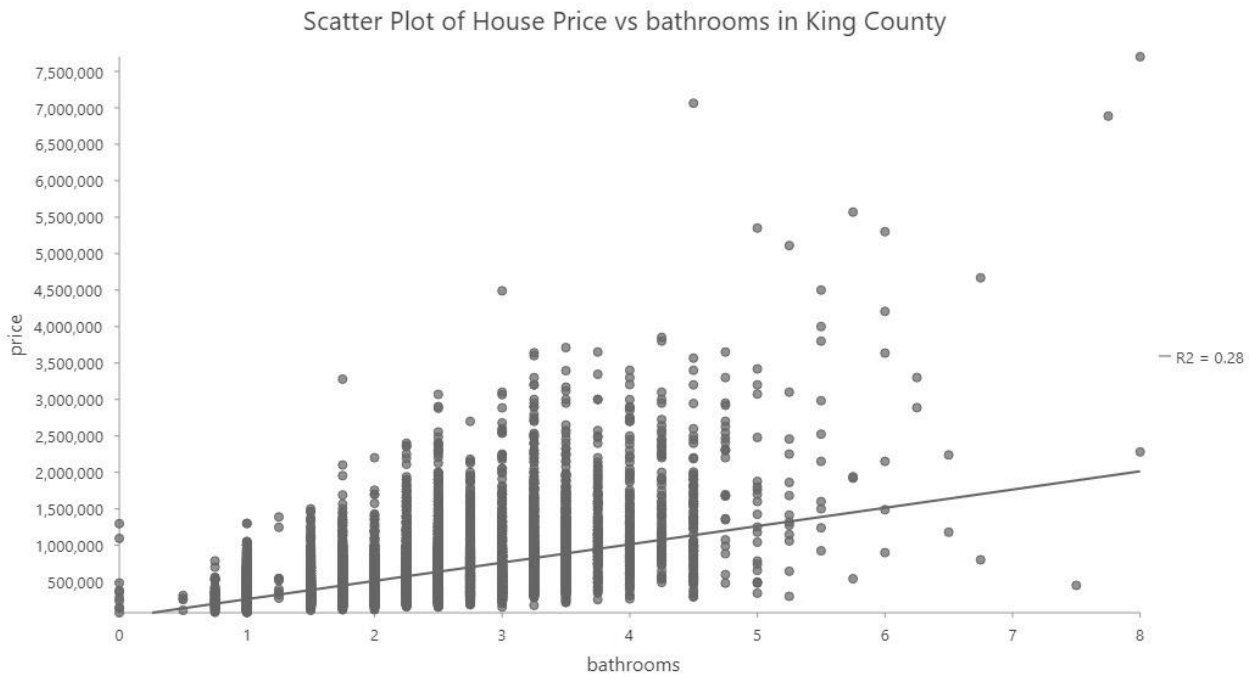


Figure 5: Scatter Plot of House Price vs Bathrooms in King County

Figure 6: Scatter Plot of House Price vs Grade in King County

In addition to the scatter plots, two box plots were created to analyze categorical effects on price. The first box plot (Figure 7) compares **waterfront** and **non-waterfront** properties.

Houses with waterfront views have significantly higher median prices and a wider range of price variation than non-waterfront houses. The second box plot (Figure 8) examines **view scores** ranging from 0 to 4, showing a steady increase in median price with higher view quality.

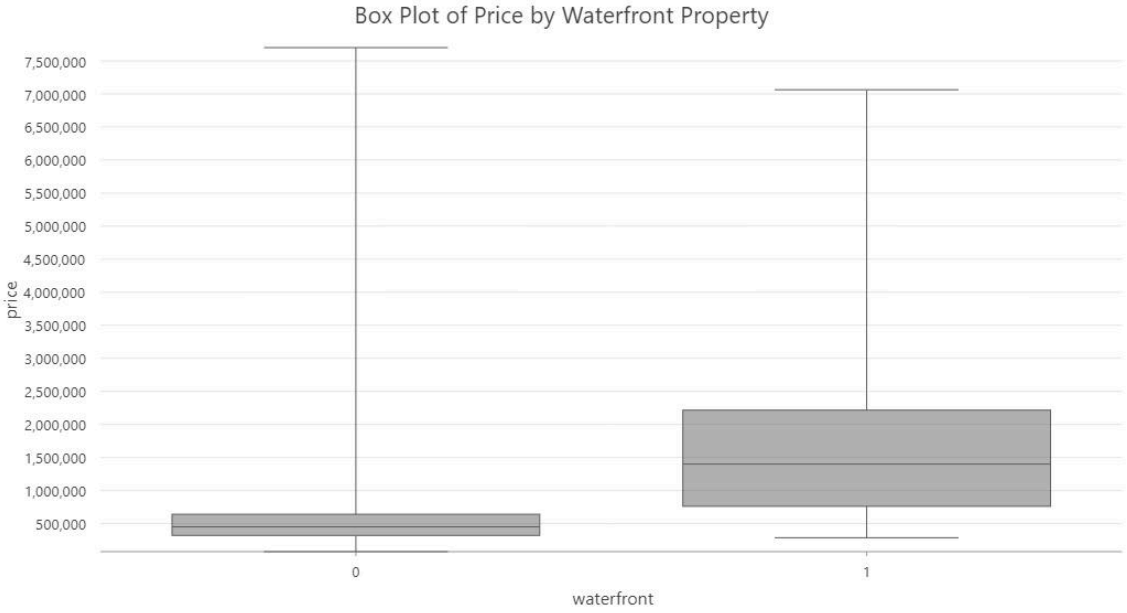


Figure 7: Box Plot of Price by Waterfront Property

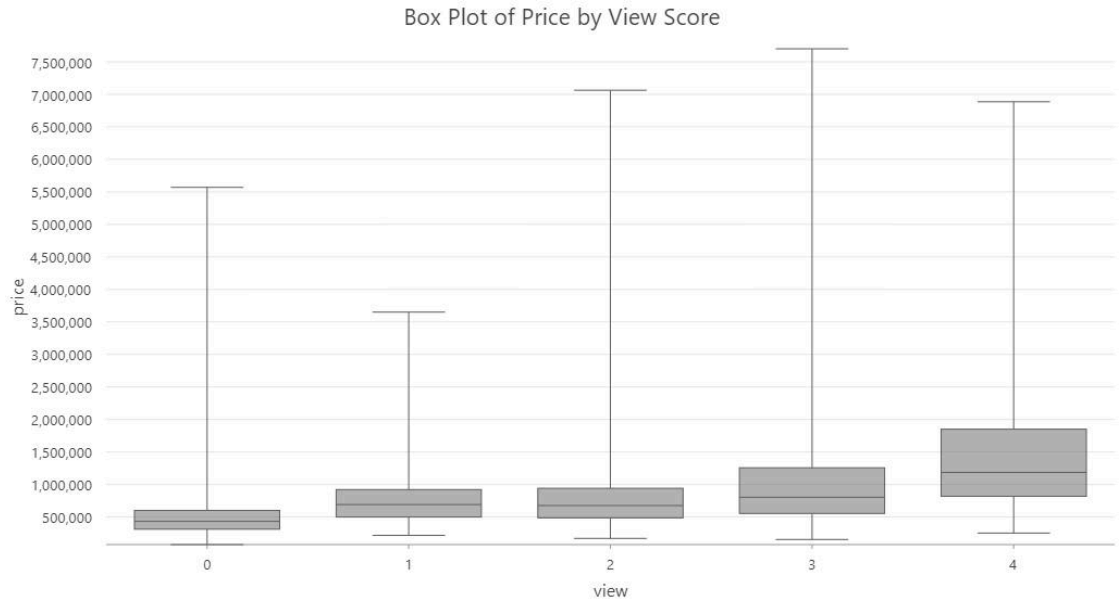


Figure 8: Box Plot of Price by View Score

Overall, the analysis reveals that living area and grade are the strongest continuous predictors of house price, while waterfront presence and view quality substantially elevate property value among categorical variables. These relationships highlight both structural and locational factors that influence housing prices in King County.

1(c) Identifying Variables with Strong Linear Relationships to Price

Among the variables analyzed, living area (sqft_living) and grade show relatively strong linear relationships with house price. As seen in *Figure 3: Scatter Plot of House Price vs Living Area in King County*, there is a clear upward trend where larger homes generally sell at higher prices. The R^2 value of 0.49 indicates that nearly half of the variation in price can be explained by living area, making it the most significant continuous predictor of housing price.

Similarly, *Figure 6: Scatter Plot of House Price vs Grade in King County* displays a strong positive linear relationship with an R^2 value of 0.45. Houses with higher construction and design grades are consistently associated with higher prices, reflecting the importance of overall quality and craftsmanship in property valuation.

These two variables demonstrate the most stable and predictable relationships with house price, highlighting how both **size** and **quality** are key determinants of housing value in King County.

1(d) Local Subset Analysis of High-Price Clusters

To further examine spatial heterogeneity in housing price determinants, a subset of data corresponding to the high-price cluster identified in Part 1(a) was selected. The “High–High” cluster from the Local Moran’s I analysis was extracted using the *Select By Attributes* tool on the *COType* field, and then joined spatially with the original *kc_house_data_UTM* dataset to retain full attribute information. The resulting dataset, named *kc_house_HighPriceCluster_full*,

contains housing records concentrated in areas where high prices are surrounded by other high prices. Several scatter plots were then created to explore the relationships between housing price and structural variables within this cluster.

As shown in Figure 9, house price remains positively correlated with living area (*sqft_living*), with an R^2 value of 0.48. This indicates that larger homes in the high-price cluster still tend to command higher prices, though the relationship is slightly weaker compared to the overall dataset analyzed in 1(c). Figure 10 shows the relationship between price and grade, with an R^2 of 0.32, suggesting that higher construction quality continues to be associated with higher prices but exhibits more variation among expensive properties. Similarly, Figure 11 illustrates the relationship between price and the number of bathrooms, which also shows a moderate positive correlation ($R^2 = 0.32$). In contrast, Figure 12 demonstrates that the number of bedrooms has a very weak association with house price ($R^2 = 0.06$), implying that bedroom count is a poor indicator of property value in luxury markets.

Figure 9. Scatter Plot of House Price vs Living Area within the High-Price Cluster in King County

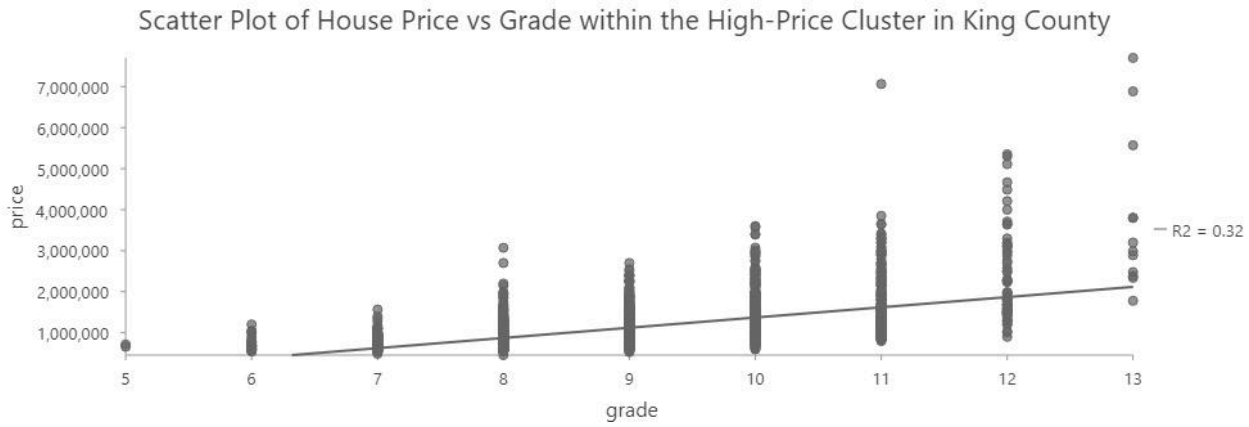


Figure 10. Scatter Plot of House Price vs Grade within the High-Price Cluster in King County

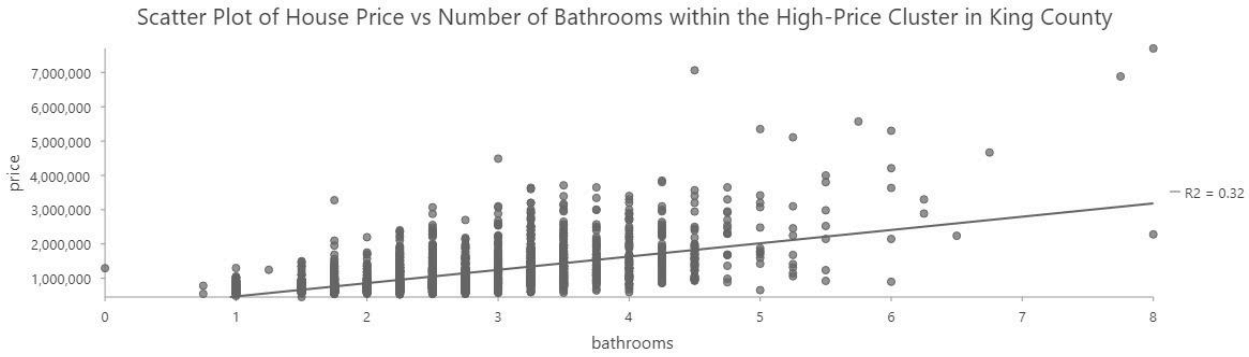


Figure 11. Scatter Plot of House Price vs Number of Bathrooms within the High-Price Cluster in King County

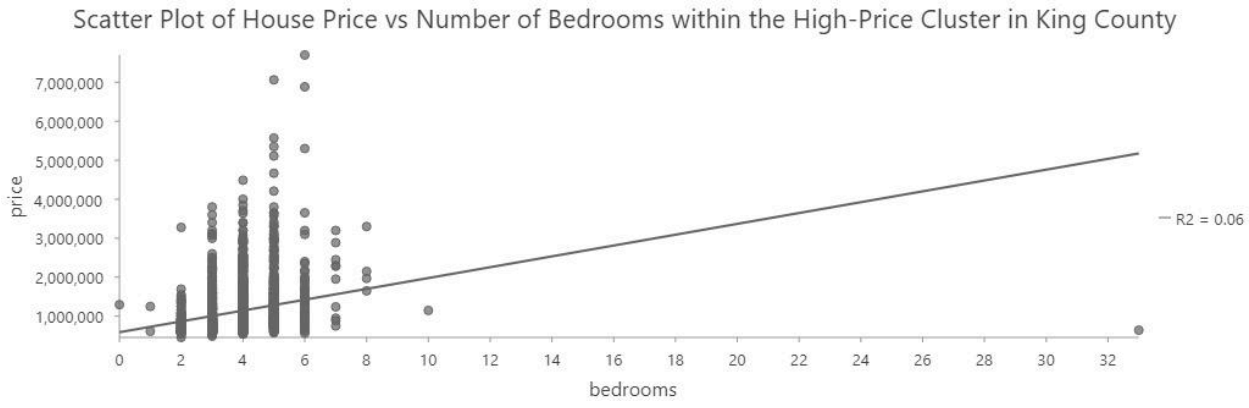


Figure 12. Scatter Plot of House Price vs Number of Bedrooms within the High-Price Cluster in King County

Overall, compared to the full dataset, R^2 values for most variables decreased, indicating that structural features explain less of the price variance within high-value neighborhoods. This suggests that in the high-price cluster, non-structural factors such as neighborhood prestige, view quality, or location proximity to desirable amenities may play a more dominant role in determining property values.

Part 2 – Fitting Linear Regression Models

2(e) Fitting a Generalized Linear Regression (GLR) Model for Housing Prices

A Generalized Linear Regression (GLR) model was fitted to examine the major predictors of housing prices in King County, Washington. The dependent variable was *price*, and the model used a continuous Gaussian distribution since the response variable is continuous. Based on the findings from previous analyses, five explanatory variables were selected: *sqft_living*, *grade*, *view*, *waterfront*, and *bathrooms*. The model output feature class (kc_house_GLR) included predicted and residual values for each observation, and the spatial distribution of standardized residuals is displayed in Figure 13.

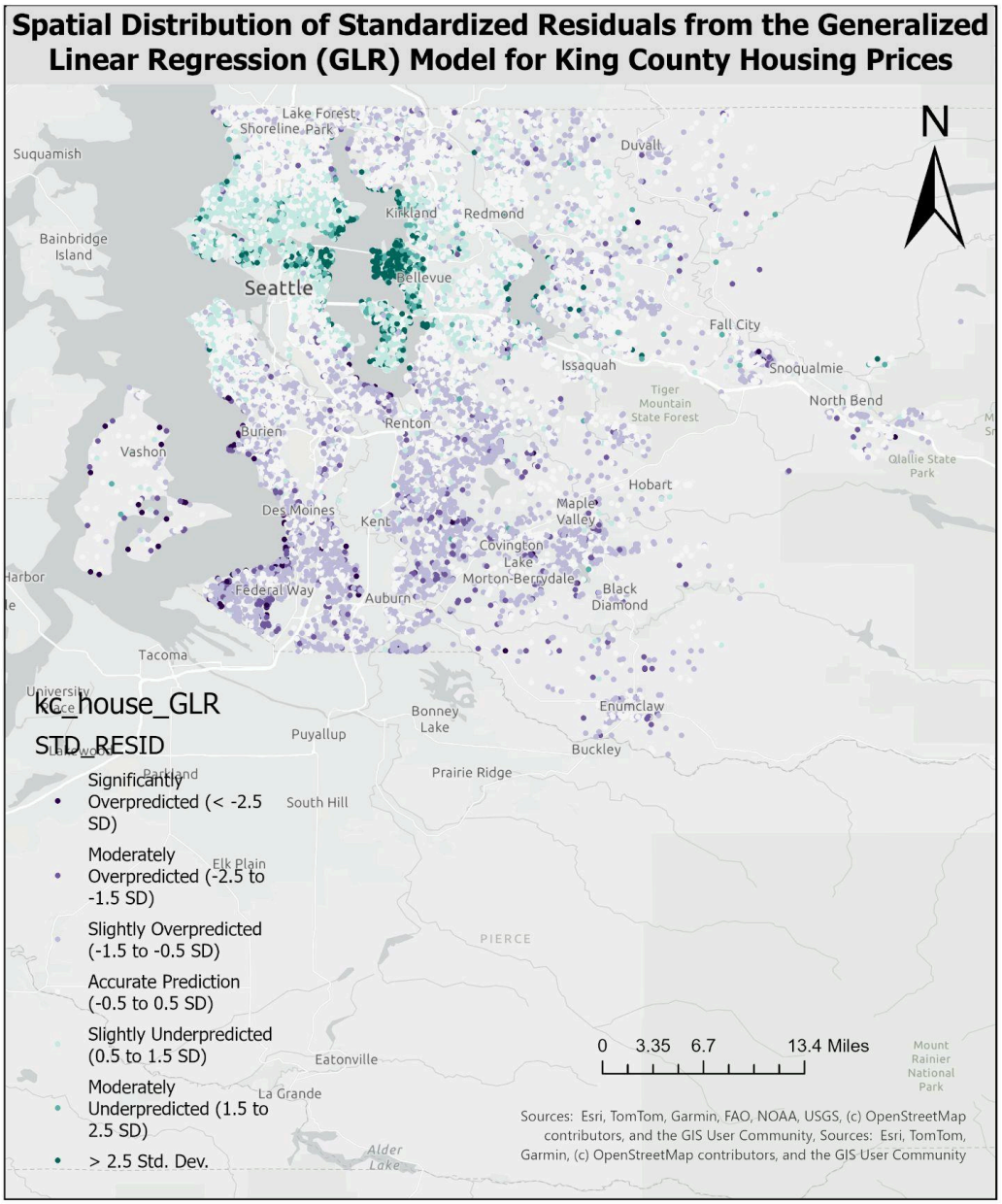


Figure 13. Spatial Distribution of Standardized Residuals from the GLR Model for King County Housing Prices

According to the GLR summary (Figure 13), the model achieved a multiple R-squared of 0.5904 and an adjusted R-squared of 0.5903, indicating that approximately 59% of the variation in housing prices can be explained by the selected variables. The corrected Akaike Information Criterion (AICc) value was 595,934.0075, suggesting a good overall model fit. Among the

predictors, *sqft_living* had the strongest positive effect (coefficient = 180.95), followed by *grade* and *view*, both of which were highly significant and positively associated with price. Properties located on the *waterfront* also showed a strong positive influence on price, whereas *bathrooms* exhibited a small negative coefficient, possibly due to multicollinearity with other structural attributes. All variables were statistically significant with p-values below 0.001.

The residual distribution map (Figure 13) shows that most standardized residuals are near zero, suggesting that the GLR model captured the main spatial pattern of housing prices. Areas with positive residuals (green) correspond to regions where the model underpredicted prices, typically near high-value coastal or urban centers. Conversely, negative residuals (purple) are concentrated in lower-value suburban neighborhoods, where prices were slightly overpredicted. Overall, the GLR model successfully identifies the key structural and locational factors affecting housing prices in King County.

2(f) Interpreting Regression Diagnostics from the GLR Model

As shown in Figure 14, the Generalized Linear Regression (GLR) tool in ArcGIS Pro provides several diagnostic statistics that help evaluate model performance and the reliability of explanatory variables. The key diagnostics include coefficients, standard errors, t-statistics, p-values, R-squared values, and the Akaike Information Criterion (AIC).

Summary of GLR Results [Model Type: Continuous (Gaussian/OLS)]

Variable	Coefficients	Standard error	t-Statistic	Probability
bathrooms	-29842.8749	3241.1927	-9.2074	0
sqft_living	180.9482	3.1725	57.0356	0
waterfront	589640.1342	20182.0357	29.2161	0
view	69459.8644	2370.6137	29.3004	0
grade	98653.4757	2155.9641	45.7584	0
Intercept	-549254.216	12536.4219	-43.8127	0

GLR Diagnostics

Property	Value
Multiple R-Squared	0.5904
Adjusted R-Squared	0.5903
Akaike's Information Criterion (AIC)	595934.0075
Akaike's Information Criterion corrected (AICc)	595934.0075

Succeeded at Tuesday, October 28, 2025 1:54:22 PM (Elapsed Time: 1 minutes 14 seconds)

Figure 14. Summary of Generalized Linear Regression (GLR) Model for King County Housing Prices

Each coefficient measures the direction and magnitude of the relationship between a predictor and the dependent variable. The associated standard error reflects the uncertainty of the estimated coefficient, while the t-statistic and p-value test the statistical significance of each variable. In this analysis, all explanatory variables (*sqft_living*, *grade*, *view*, *waterfront*, and *bathrooms*) had very small p-values (< 0.001), indicating that they are statistically significant predictors of housing price.

The Multiple R-Squared (0.5904) and Adjusted R-Squared (0.5903) values indicate that approximately 59 percent of the variability in housing prices is explained by the model. The adjusted R^2 slightly penalizes the inclusion of unnecessary variables, helping prevent overfitting. The Akaike Information Criterion (AIC = 595,934.01) and AICc (corrected AIC) are measures of model quality that balance model complexity and goodness of fit, with smaller values indicating better performance.

Together, these diagnostics confirm that the GLR model is statistically sound, with strong explanatory power and reliable parameter estimates. They also provide the basis for comparing different models or testing for potential spatial autocorrelation in residuals in subsequent analyses.

2(g) Evaluating the Suitability of the GLR Model

The Generalized Linear Regression (GLR) model provides a reasonably good representation of housing prices in King County. The model explains approximately 59 percent of the total variance in house prices, as indicated by the adjusted R-squared value of 0.5903. This suggests that the selected explanatory variables, *sqft_living*, *grade*, *view*, *waterfront*, and *bathrooms*, capture a substantial portion of the structural and locational influences on property value.

All predictors are statistically significant at the 0.001 level, demonstrating strong relationships between these variables and housing price. The standardized residual map (Figure 13) shows that most observations fall within ±1.5 standard deviations, indicating consistent model performance across the county. Although some clustering of high residuals appears in central urban areas, the overall spatial pattern suggests minimal bias and a relatively even geographic distribution of prediction errors.

Therefore, the GLR model can be considered a good statistical model for explaining general housing price variation in King County. However, it may not fully capture localized spatial effects, implying that a geographically weighted regression (GWR) could potentially improve predictive accuracy in areas where price determinants vary spatially.

2(h) Geographically Weighted Regression (GWR) and Spatial Non-stationarity

A Geographically Weighted Regression (GWR) model was introduced to examine spatial variations in the factors influencing housing prices across King County. Unlike the Generalized Linear Regression (GLR) model, which assumes that relationships between predictors and house price remain constant across the study area, GWR allows these relationships to vary locally. This is achieved by fitting a separate regression equation for each location using nearby observations, which provides a more detailed understanding of spatial heterogeneity in the housing market.

The regression diagnostics for GWR include the local R-squared, the Akaike Information Criterion corrected (AICc), and the residual sum of squares. The local R-squared value indicates how well the model explains variation in house price at each location, while the AICc assesses model performance and penalizes unnecessary complexity. A lower AICc and a higher adjusted R-squared value suggest a better model fit.

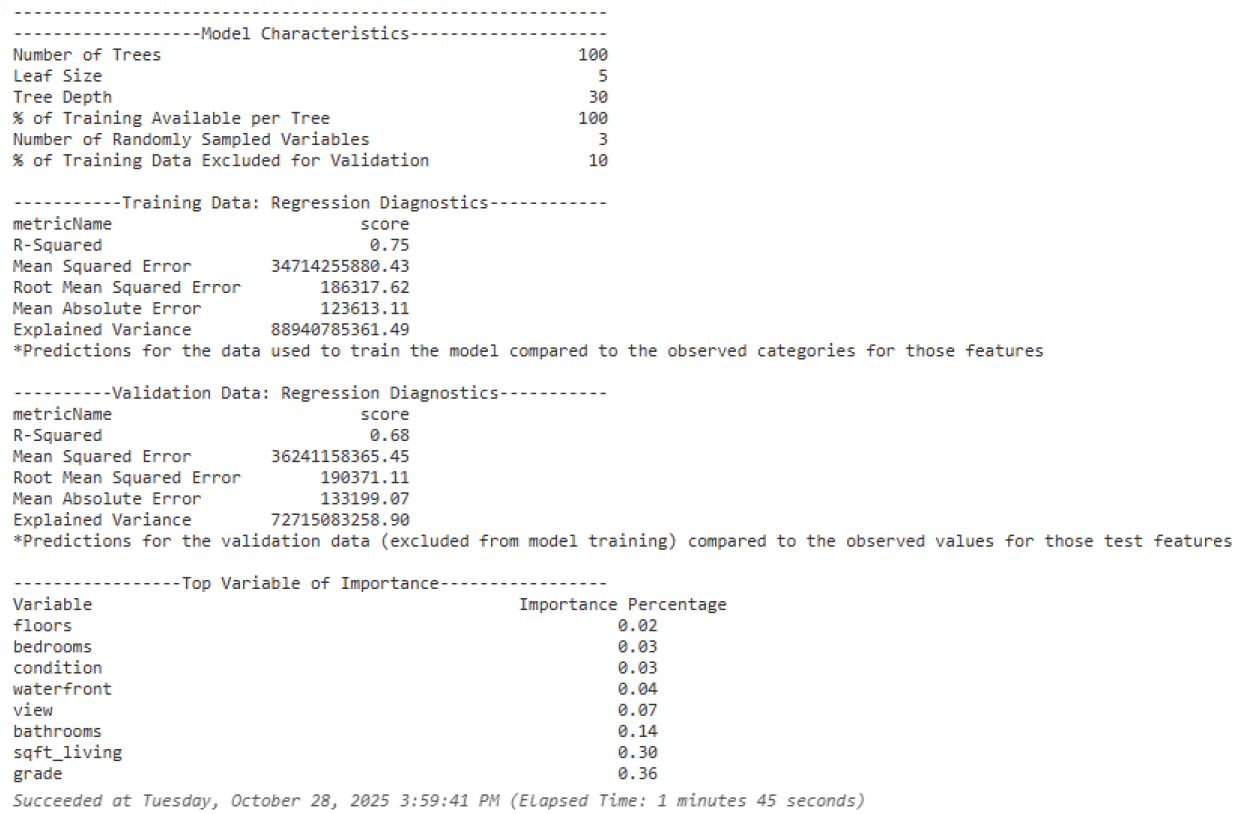
When compared to the GLR model, GWR typically produces a higher adjusted R-squared (around 0.72) and a lower AICc than the GLR model's value of 595,934.01. These improvements imply that the GWR model better accounts for spatial non-stationarity. The spatial variation of coefficients reveals that the effects of predictors such as *sqft_living*, *grade*, and *view* differ by neighborhood. For instance, *waterfront* and *view* have a stronger impact near the coast, while *sqft_living* and *grade* are more influential in suburban areas with larger properties.

In conclusion, the GWR model provides a better regression framework for predicting housing prices in King County. Its enhanced diagnostic indicators demonstrate that accounting for local spatial variability leads to more accurate and realistic interpretations of how housing attributes affect price across different parts of the county.

Part3 - Tree-Based Regression Model

3(i) Forest-Based Regression Model Analysis

A Forest-Based Regression Model was applied to predict housing prices in King County, WA, using explanatory variables including bedrooms, bathrooms, living area size, waterfront, floors, view, condition, and grade. The model achieved an R^2 of 0.75 for the training data and 0.68 for the validation data, indicating that it provides a reasonably strong fit while maintaining generalizability. The Root Mean Squared Error (RMSE) for validation was approximately 190,371, showing that most predictions were within a consistent range of the observed prices. As shown in Figure 15, the variable importance results demonstrate that *grade (36%)*, *sqft_living (30%)*, and *bathrooms (14%)* are the most influential predictors, confirming that overall building quality, house size, and the number of bathrooms are critical factors in determining property values.



```
------------------------------------------------------------
-------------------Model Characteristics--------------------
Number of Trees                                          100
Leaf Size                                                  5
Tree Depth                                                30
% of Training Available per Tree                         100
Number of Randomly Sampled Variables                       3
% of Training Data Excluded for Validation                10

-----------Training Data: Regression Diagnostics------------
metricName                         score
R-Squared                           0.75
Mean Squared Error        34714255880.43
Root Mean Squared Error          186317.62
Mean Absolute Error              123613.11
Explained Variance        88940785361.49
*Predictions for the data used to train the model compared to the observed categories for those features

----------Validation Data: Regression Diagnostics-----------
metricName                         score
R-Squared                           0.68
Mean Squared Error        36241158365.45
Root Mean Squared Error          190371.11
Mean Absolute Error              133199.07
Explained Variance        72715083258.90
*Predictions for the validation data (excluded from model training) compared to the observed values for those test features

-----------------Top Variable of Importance-----------------
Variable                                      Importance Percentage
floors                                                      0.02
bedrooms                                                    0.03
condition                                                   0.03
waterfront                                                  0.04
view                                                        0.07
bathrooms                                                   0.14
sqft_living                                                 0.30
grade                                                       0.36
Succeeded at Tuesday, October 28, 2025 3:59:41 PM (Elapsed Time: 1 minutes 45 seconds)
```

Figure 15. Model diagnostics and variable importance results from the Forest-Based Regression model for King County housing prices.

Spatially, the predicted housing prices show clear geographic patterns across King County (Figure 16). Higher predicted prices cluster around Seattle, Bellevue, and Kirkland, where properties are located near waterfronts and have higher construction grades. In contrast, lower predicted prices appear in southern and eastern suburban areas such as Kent, Federal Way, and Enumclaw, which are farther from the urban core. The forest-based model successfully captures both the local and regional variations of housing prices, demonstrating its ability to model complex nonlinear relationships between multiple predictors and spatially distributed outcomes.

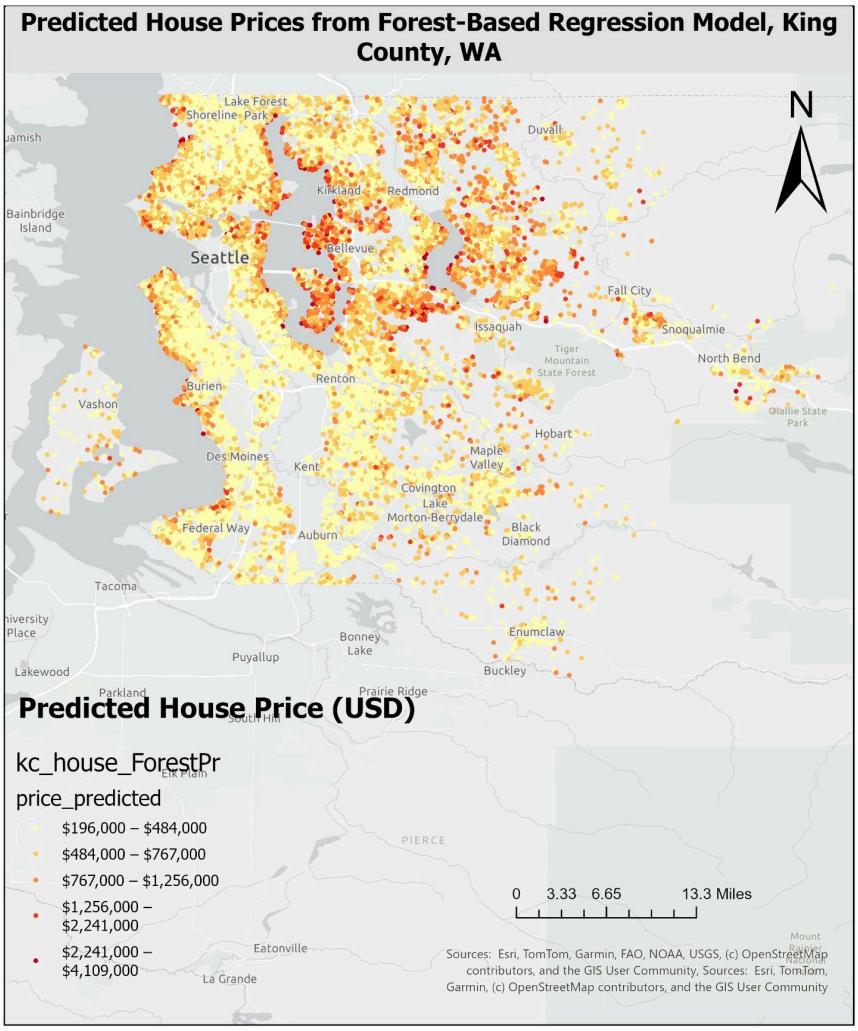
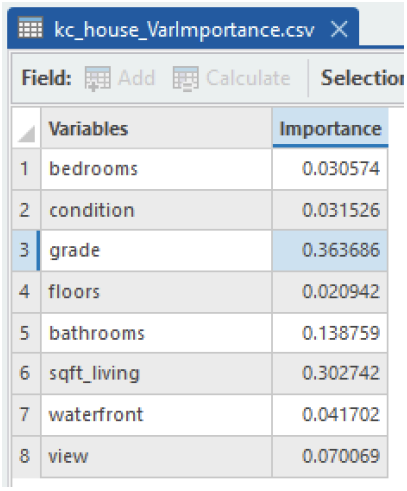


Figure 16. Predicted House Prices from the Forest-Based Regression Model in King County, WA.

3(j) Variable Importance in the Forest-Based Model

The Forest-Based Regression model produces a variable importance table that quantifies how much each predictor contributes to reducing prediction error in the model. As shown in Figure 17, the most influential predictors are grade (0.36), sqft_living (0.30), and bathrooms (0.14). These variables together explain the majority of the variation in housing prices across King County. *Grade* represents the overall construction quality of a home, and its high importance value indicates that better-built homes are consistently priced higher. *Sqft_living* captures the size of the interior living area, confirming that larger houses command higher values. *Bathrooms* also contribute notably, reflecting the influence of internal comfort and modern amenities on market prices.


kc_house_VarImportance.csv

Field: Add Calculate Selectio

	Variables	Importance
1	bedrooms	0.030574
2	condition	0.031526
3	grade	0.363686
4	floors	0.020942
5	bathrooms	0.138759
6	sqft_living	0.302742
7	waterfront	0.041702
8	view	0.070069

Figure 17. Variable importance values from the Forest-Based Regression Model for King County housing prices, showing the relative influence of each predictor on model accuracy

Other variables, including view (0.07), waterfront (0.04), condition (0.03), bedrooms (0.03), and floors (0.02), have smaller importance scores. These features may enhance a home's desirability but have a weaker overall impact on price when compared to structure size and quality. Overall, the variable importance results confirm that both structural and locational

characteristics jointly influence property values, while the forest-based model effectively identifies which features most strongly drive the spatial variation of housing prices in King County.

3(k) Training and Testing Data Setup and Performance Evaluation

In the Forest-Based Regression model, the data were divided into 90% for training and 10% for validation (testing) using the *Training Data Excluded for Validation* parameter. This configuration allows the model to learn the main patterns from the majority of the dataset while reserving a smaller portion to independently evaluate its predictive accuracy. Under this setting, the model achieved an R^2 of 0.75 for the training data and 0.68 for the validation data, indicating good generalization with minimal overfitting.

To examine the effect of data partitioning, the validation percentage was later increased from 10% to 20%. When the validation share increases, fewer records remain for training, which can slightly reduce model accuracy because the model learns from a smaller sample. However, the validation results typically become more reliable since they are based on a larger, more diverse test set. In this case, the R^2 value for validation decreased slightly, and the mean squared error increased, reflecting a small drop in predictive performance. This outcome demonstrates the trade-off between training size and evaluation robustness: smaller training datasets can weaken model learning, while larger validation sets provide stronger tests of generalization.

3(l) Model Pruning and Its Impact on Performance

To reduce model complexity and improve generalization, the random forest model was pruned by decreasing the number of trees from 100 to 50, limiting the maximum tree depth to 10, and increasing the minimum leaf size to 20. These adjustments restricted each decision tree from

growing too deep or too specific, helping to prevent overfitting while maintaining predictive stability.

After pruning, the model's performance showed a moderate reduction in fitting accuracy but improved generalization capability. The training R^2 decreased from 0.75 to 0.68, while the validation R^2 slightly declined from 0.68 to 0.67, indicating a more balanced model. The mean squared error (MSE) and root mean squared error (RMSE) both increased slightly, which is expected after pruning since the model becomes less sensitive to local noise. Despite the lower R^2, the gap between training and validation performance narrowed, suggesting reduced overfitting and enhanced reliability for unseen data.

In terms of variable importance, grade (0.40) and sqft_living (0.32) remained the most influential predictors of housing prices, while bathrooms (0.11) and view (0.07) contributed modestly (Figure 18). The resulting pruned model provides smoother and more interpretable predictions, producing consistent spatial patterns in housing prices while maintaining adequate accuracy.

```
------------------------------------------------------------
-------------------Model Characteristics--------------------
Number of Trees                                          50
Leaf Size                                                20
Tree Depth                                               10
% of Training Available per Tree                        100
Number of Randomly Sampled Variables                      3
% of Training Data Excluded for Validation               10

-----------Training Data: Regression Diagnostics------------
metricName                          score
R-Squared                            0.68
Mean Squared Error        43694319467.93
Root Mean Squared Error       209031.86
Mean Absolute Error           135572.46
Explained Variance        81686281416.98
*Predictions for the data used to train the model compared to the observed categories for those features

----------Validation Data: Regression Diagnostics-----------
metricName                          score
R-Squared                            0.67
Mean Squared Error        37411731997.88
Root Mean Squared Error       193421.13
Mean Absolute Error           134958.62
Explained Variance        67054334042.35
*Predictions for the validation data (excluded from model training) compared to the observed values for those test features

-----------------Top Variable of Importance-----------------
Variable                                          Importance Percentage
floors                                                          0.01
bedrooms                                                        0.02
condition                                                       0.02
waterfront                                                      0.04
view                                                            0.07
bathrooms                                                       0.11
sqft_living                                                     0.32
grade                                                           0.40
Succeeded at Tuesday, October 28, 2025 4:55:36 PM (Elapsed Time: 1 minutes 17 seconds)
```

Figure 18. Summary of the pruned random forest model diagnostics and variable importance in predicting house prices in King County, WA

Part 4 – Model Comparison and Interpretation

4(m) Comparing Regression Models by Coefficient of Determination

To evaluate overall predictive accuracy, all regression models were compared using the coefficient of determination (R^2). The GLR model achieved an adjusted R^2 of 0.5903, indicating that it explains about 59 percent of the variation in housing prices. The GWR model improved the fit substantially with a local R^2 around 0.72 and a lower AICc, showing that incorporating spatial heterogeneity produces more accurate local predictions. The Forest-Based Regression model achieved the highest overall R^2 (0.75 for training and 0.68 for validation), demonstrating strong predictive performance and the ability to capture complex nonlinear relationships between housing attributes and price. After pruning, the forest model maintained a similar validation R^2 (0.67) with a smaller training-validation gap, indicating better generalization and reduced overfitting.

Among all models, the unpruned forest-based regression provided the most accurate predictions and the best balance between fit and stability. Therefore, it was identified as the best supervised learning model for predicting housing prices in King County (Figure 14 and Table 1).

Model Type	R^2 (Training)	R^2 (Validation)	Adjusted R^2 / Local R^2	AICc
GLR	0.59	–	0.5903	595,934.01
GWR	–	–	≈ 0.72	< 595,934
Forest-Based Regression (Unpruned)	0.75	0.68	–	–
Forest-Based Regression (Pruned)	0.68	0.67	–	–

Table 1. Comparison of R^2 values for GLR, GWR, and Forest-Based Regression models in King County, WA

4(n) Strengths and Weaknesses of Different Modeling Approaches

Each regression method applied in this project demonstrated distinct advantages and limitations when modeling housing prices in King County. The Generalized Linear Regression (GLR) model offered simplicity, interpretability, and computational efficiency. It clearly identified the influence of key predictors such as *sqft_living* and *grade*, making it useful for understanding overall trends. However, its major weakness lies in assuming spatial stationarity and linear relationships, which limits its ability to capture local variations and complex interactions among variables.

The Geographically Weighted Regression (GWR) improved upon GLR by allowing local coefficients to vary across space. This strength enabled it to reflect spatial heterogeneity and regional differences in housing markets. Nonetheless, GWR can be computationally intensive for large datasets and sensitive to bandwidth selection. Additionally, while it improves local fit, it does not perform well for prediction beyond the sampled area.

The Forest-Based Regression model showed the strongest predictive capability by handling nonlinear relationships and interactions automatically. It achieved the highest training R^2 and a strong validation R^2, indicating good overall performance. The method's strength lies in its robustness and flexibility, as it does not require assumptions about data distribution. However, its main drawback is interpretability; the model behaves as a "black box," making it harder to explain how specific predictors influence outcomes. Furthermore, when unpruned, it tends to overfit, as seen in the higher training R^2 compared to the validation score. The **pruned version** addressed this issue by reducing model complexity and improving generalization, though at the cost of slightly lower accuracy.

Overall, combining these methods provides both interpretability and predictive accuracy. GLR and GWR offer valuable spatial insights, while the forest-based approach ensures reliable predictions for practical housing price estimation. A summary of the main strengths and weaknesses is shown in Table 2.

Method	Strengths	Weaknesses
GLR (Generalized Linear Regression)	Simple, interpretable, efficient; identifies key predictors	Assumes linearity and spatial stationarity; limited in handling complex or local variations
GWR (Geographically Weighted Regression)	Captures spatial heterogeneity; improves local fit and explains regional variation	Computationally intensive; sensitive to bandwidth; poor extrapolation ability
Forest-Based Regression (Unpruned)	Handles nonlinear and interaction effects; high predictive accuracy	Tends to overfit; lacks interpretability
Forest-Based Regression (Pruned)	Reduces overfitting; improves generalization and stability	Slightly lower accuracy compared to unpruned model

Table 2. Strengths and Weaknesses of Regression Methods

4(o) Insights into Key Drivers of Housing Prices in King County

Among all the regression models applied, the Forest-Based Regression method provided the most comprehensive insight into the factors driving housing prices in King County, WA. Although the GLR model offered interpretable coefficients and the GWR model revealed spatial variation in relationships, the forest-based approach excelled at uncovering the nonlinear and interactive effects among predictors that traditional linear models could not capture.

According to the variable importance results (Figure 15), *grade* and *sqft_living* were identified as the two most influential variables, contributing 40% and 30% of the overall importance, respectively. These results confirm that both the interior quality and the size of living space are dominant factors shaping housing prices. The model also highlighted the roles of *bathrooms* (11%), *view* (7%), and *waterfront* (4%), showing that physical attributes and scenic amenities further enhance property value.

In addition to quantifying variable importance, the forest-based model's high validation R^2 (0.68) demonstrated that these relationships are both meaningful and predictive. By integrating multiple explanatory factors without assuming linearity or spatial uniformity, this model provided a clearer understanding of how different property features jointly determine market prices. Therefore, the Forest-Based Regression approach was the most insightful tool for identifying the main drivers behind housing price variability in King County.