

Multiple Linear Regression Predictive Modeling for House Prices in King County

By Sam Odongo

Overview

This project aims to develop a multiple linear regression model to predict house prices in King County.

Accurate price estimation is crucial for informed decision-making in the real estate market.

Leveraging the King County House Sales dataset, we will identify key factors impacting house prices and build a reliable regression model that can accurately predict house prices.





Business Understanding

Why this Model?

Lack of an efficient method to accurately predict house prices in King County.

Existing methods may overlook relevant features, leading to inaccurate estimations.

There is a need to develop a predictive model that considers multiple variables and accurately predicts house prices.

By using multiple predictor variables, we can capture more nuanced relationships and potentially improve the accuracy of the predictions



Main Objective

Develop an accurate predictive model for house prices in King County based on various features.

Analyze King County House Sales dataset, implement multiple regression model, and provide a reliable tool for buyers, sellers, and real estate professionals.



Specific Objectives

Develop and evaluate multiple linear regression model

Identify influential factors impacting house prices in the King County housing market.

Explore regression algorithms and feature selection techniques to determine key predictors of house prices.





Data

Data used consists of 20 columns and 21597 rows of King county house sales data that is essential for this project,

Variable	Description
id	A unique identifier for each house
date	The date when the house was sold
price	The target variable representing the price of the house
bedrooms	The number of bedrooms in the house
bathrooms	The number of bathrooms in the house
sqft_living	The square footage of the home
sqft_lot	The square footage of the lot
floors	The total number of floors in the house
waterfront	Indicates whether the house has a view to a waterfront
view	Indicates whether the house has been viewed
condition	Represents the overall condition of the house
grade	Represents the overall grade given to the housing unit
sqft_above	The square footage of the house apart from the basement
sqft_basement	The square footage of the basement
yr_built	The year the house was built
yr_renovated	The year when the house was renovated
zipcode	The zip code of the house's location
lat	The latitude coordinate of the house's location
long	The longitude coordinate of the house's location
sqft_living15	The square footage of interior housing living space for the nearest 15 neighbors
sqft_lot15	The square footage of the land lots of the nearest 15 neighbors

Data Preparation

The modeling process begins with a thorough understanding and exploration of the data.

The dataset contains various variables that provide valuable insights into house prices, including the number of bedrooms, bathrooms, square footage, condition, and other important factors.

The data was checked for missing values and handled appropriately by imputing missing values and removing rows with missing data.

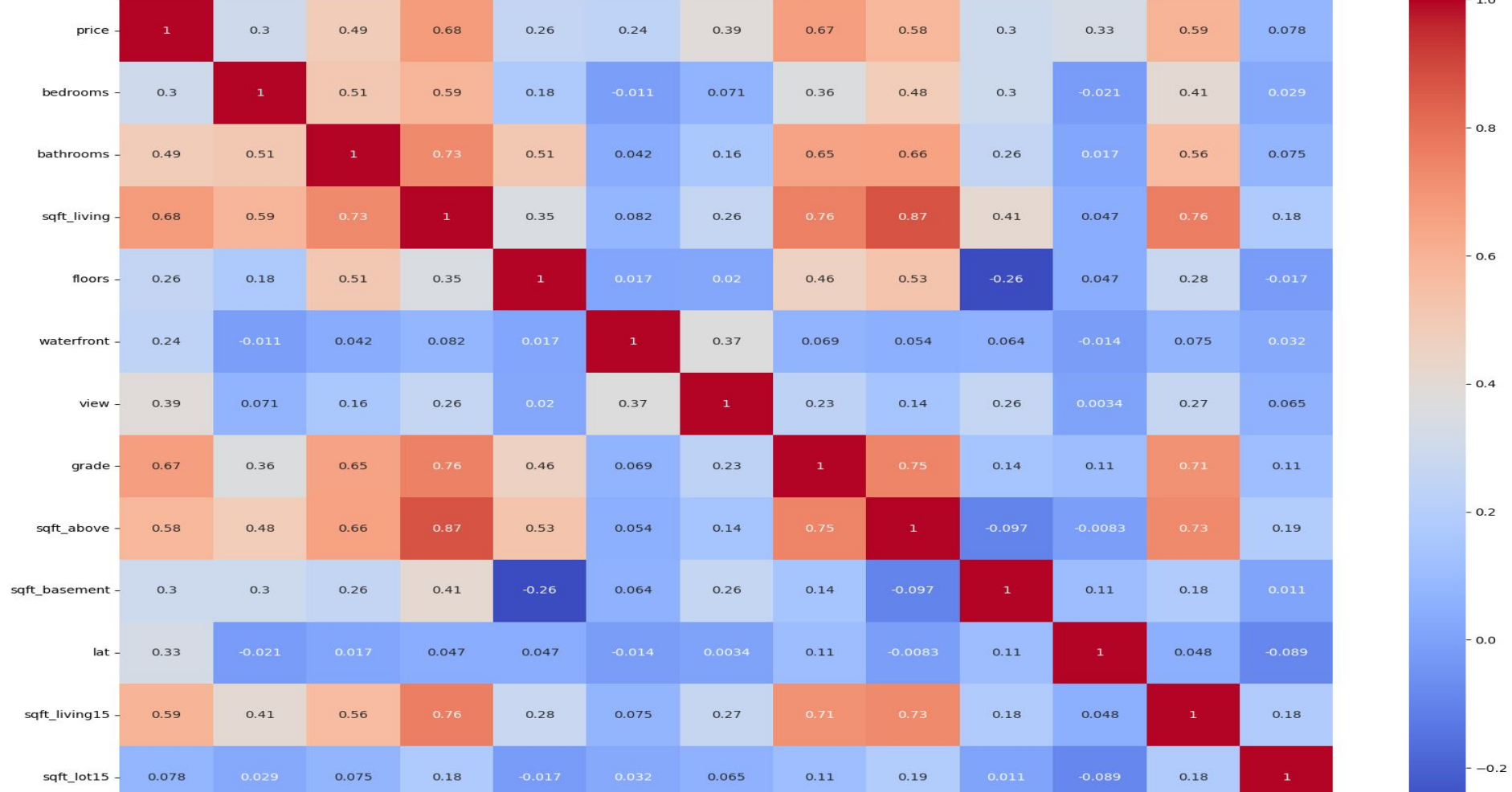
Outliers or erroneous data points that could affect the analysis and model performance were identified and handled.

Correlation: Data refined further by identifying the most promising predictors; Correlation.



Correlation between Price and the most Influential Predictors

Correlation Heatmap



Modelling

Model Formula

In the context of this project, the multiple regression formula would be:


$$\text{Price} = \beta_0 + \beta_1 \text{Grade} + \beta_2 \text{Bathrooms} + \beta_3 \text{Sqft_living15} + \beta_4 \text{Sqft_living} + \epsilon$$

β_0 is the intercept term, representing the value of Y when all predictor variables are zero.

$\beta_1, \beta_2, \dots, \beta_n$ are the regression coefficients, representing the change in Y associated with a one-unit change in each respective predictor variable.

ϵ is the error term, representing the random variability or unexplained part of Y not accounted for by the predictor variables.

The goal of the multiple regression analysis is to estimate the regression coefficients ($\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$) that provide the best fit to the data, enabling accurate predictions of house prices based on the selected predictor variables.



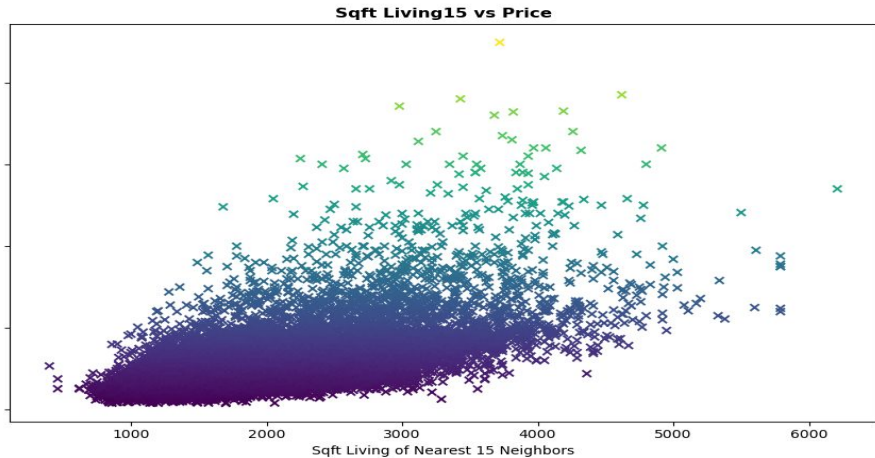
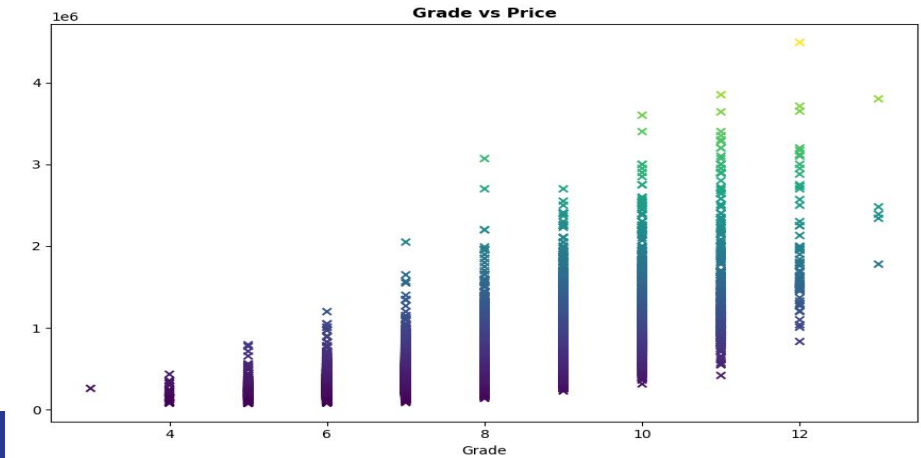
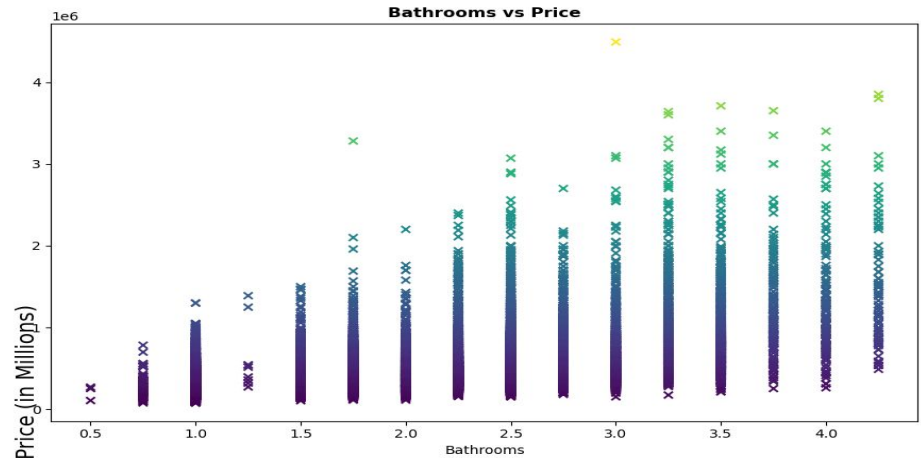
Feature Selection

For this Model I selected 4 features that had the highest correlation with price and did not violate the multicorrelianity assumption:

- grade
- bathrooms
- sqft_living15
- sqft_living



Relationship of Selected Variables vs Prices



Regression Results

OLS Regression Results

```

=====
Dep. Variable:          price    R-squared:                0.534
Model:                  OLS      Adj. R-squared:           0.534
Method:                 Least Squares    F-statistic:             6000.
Date:                   Fri, 07 Jul 2023    Prob (F-statistic):       0.00
Time:                   00:34:39    Log-Likelihood:          -7786.2
No. Observations:       20913    AIC:                     1.558e+04
Df Residuals:           20908    BIC:                     1.562e+04
Df Model:                4
Covariance Type:        nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	8.1050	0.079	102.451	0.000	7.950	8.260
bathrooms	-0.0228	0.005	-4.350	0.000	-0.033	-0.013
grade	0.1879	0.003	55.837	0.000	0.181	0.194
sqft_living	0.3621	0.011	31.856	0.000	0.340	0.384
sqft_living15	0.1830	0.012	15.536	0.000	0.160	0.206

```

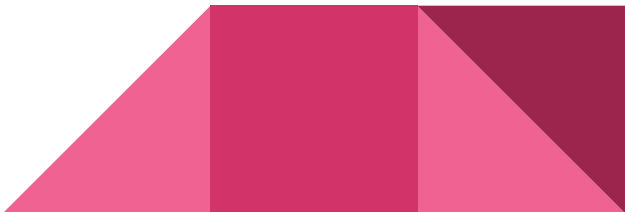
=====
Omnibus:                80.121    Durbin-Watson:           1.977
Prob(Omnibus):           0.000    Jarque-Bera (JB):        75.727
Skew:                    0.120    Prob(JB):                3.60e-17
Kurtosis:                2.830    Cond. No.                 390.
=====

```


Model validation:MSE AND RMSE

Train Root Mean Squared Error	0.34961944187687516
Test Root Mean Squared Error	0.35565383384411436

Train Mean Squared Error	0.1222337541382977
Test Mean Squared Error	0.1264896495280169



Is the Model Successful?

The regression model is linear in nature

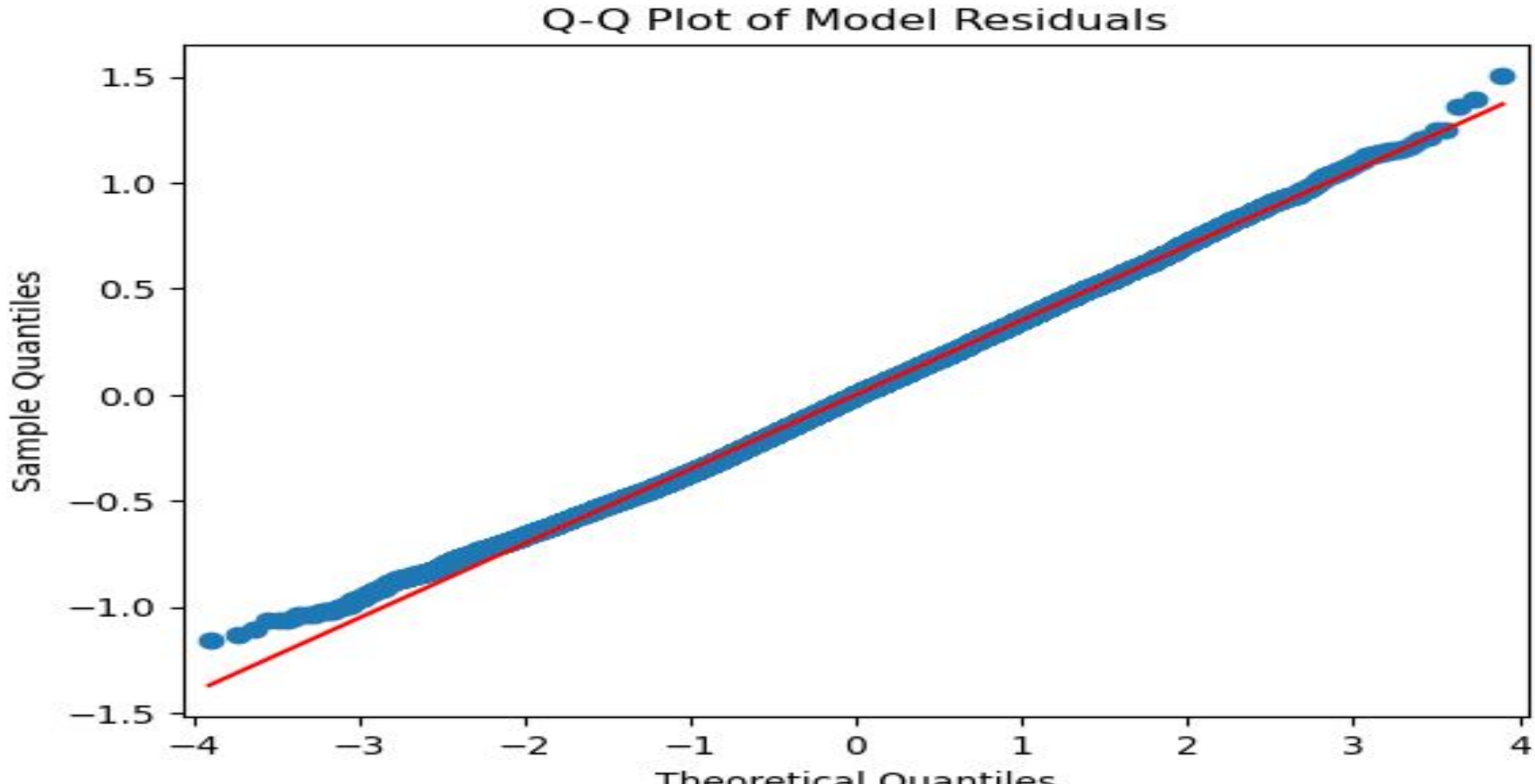
The errors are independent

The error terms are normally distributed

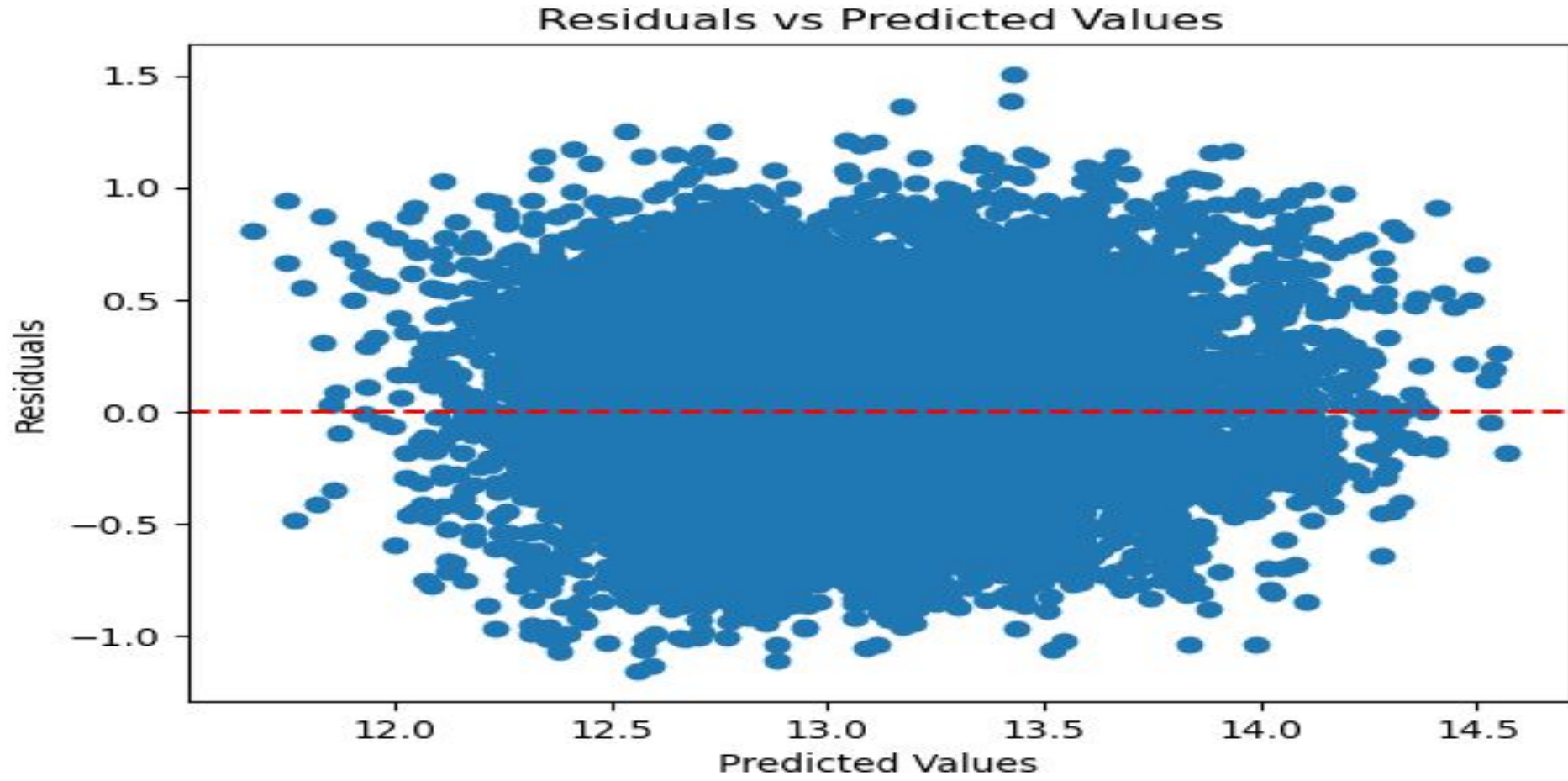
The error has a constant variance



Predicted vs Actual Price



Residuals vs Predicted Values



Interpretation

The R-squared value of 53.4% indicates that approximately 53.4% of the variation in house prices can be explained by the included variables

The coefficients :provide valuable insights into the impact of each variable on house prices.

- **Bathrooms:** Each additional bathroom is associated with a decrease of \$22,800 in house price, holding other factors constant.
- **Grade:** A higher grade is linked to a higher house price, with each unit increase in grade leading to an increase of \$187,900 in price.
- **Square Footage of Living Space:** An increase in square footage of living space is associated with a higher house price, with each unit increase resulting in an increase of \$362,100 in price.
- **Square Footage of Interior Housing Living Space for the Nearest 15 Neighbors (sqft_living15):** More living space among the nearest 15 neighbors is connected to a higher house price, with each unit increase leading to an increase of \$183,000 in price.

All coefficients have low p-values, indicating their statistical significance in influencing house prices.



Conclusions

The multiple regression model successfully predicts house prices in the King County housing market, explaining approximately 53.4% of the price variation. The model's assumptions, such as linearity, are reasonably satisfied.

The model demonstrates the significance of bathrooms, grade, square footage of living space, and square footage of interior housing living space for the nearest 15 neighbors in determining house prices. Each variable has a significant impact, and their coefficients provide actionable insights for buyers, sellers, and real estate professionals.

The low MSE and RMSE values validate the model's accuracy in estimating house prices. The consistency in performance across the train and test sets further strengthens the model's validity.



Recommendations

Consumers and investors should consider the number of bathrooms, grade, square footage of living space, and square footage of interior housing living space for the nearest 15 neighbors when evaluating house prices.

Pay attention to the quality and grade of a house, as higher grades are associated with higher prices.

Consider the impact of square footage of living space on house prices. A larger living space is associated with higher prices.

Individual preferences regarding bathrooms should be considered alongside other factors.

Data professionals should validate the model's performance regularly using additional evaluation metrics.

Explore alternative modeling approaches and additional variables to further enhance the model's predictive power.



Future Work

Explore alternative modeling techniques such as decision trees, random forests, or gradient boosting to potentially improve predictive performance.

Consider incorporating additional variables that may impact house prices, such as location factors, amenities, or economic indicators.

Conduct further analysis to assess potential interactions or non-linear relationships between variables.

Validate the model with updated data to ensure its relevance and accuracy in a changing real estate market.

