The Real Estate

Analyzing Price House Sale Influencers in King County.

End User:

Clients:

1. Seller, a resident of King County looking to sell their house for the best price.

2. Buyer, a random person looking to buy a house in King County at an affordable price.

Report By:

Bonventure Osoro Willis.

Introduction.

our project is focused on assisting both buyers and sellers in the dynamic real estate market of King County. We understand the unique challenges and opportunities that arise for both parties during the process of buying or selling a property. Our objective is to provide valuable insights and recommendations to help buyers make informed decisions and find their dream homes, while also guiding sellers to optimize their sales strategies and achieve the best possible prices for their properties. Through our analysis of house sales in King County, we aim to empower both buyers and sellers with the knowledge they need to navigate the market successfully.

Report Overview.

1. Business Understanding
2. Data Assembly And Preparation
3. Modelling
4. Regression Results
5. Model Refinement and Evaluation
6. Summary.
7. Recommendations.

Problem Statement

The goal of this project is to analyze the factors influencing house sale prices in King County and provide actionable insights for both buyers and sellers. By understanding the key predictors, buyers can make informed investment decisions and sellers can set competitive listing prices. The analysis aims to bridge the information gap in the real estate market and empower clients with data-driven recommendations. Through this project, we will identify significant predictors and provide tailored guidance to thrive in the dynamic and competitive real estate landscape of King County.

Objective.

Based on the problem statement, the objectives of the project are to identify significant factors influencing house sale prices, create accurate linear regression models to predict sale prices, analyze the models' performance, provide quantifiable recommendations to buyers based on the analysis of significant factors, and provide quantifiable recommendations to sellers to optimize their pricing strategies. By achieving these objectives, the project aims to empower both buyers and sellers in making informed decisions in the competitive real estate market of King County.

Metrics For Success

The success of the project will be evaluated based on achieving a high R-squared and adjusted R-squared value, indicating a strong fit of the regression models to the data and a high proportion of variance explained. Additionally, a significant F-statistic with a low p-value will demonstrate the overall significance of the models, while a Durbin-Watson statistic close to 2 will indicate no significant residual autocorrelation. Finally, significant coefficients with low p-values will validate the impact of individual predictors on house prices. By meeting these metrics, the project will provide accurate predictions and actionable insights for buyers and sellers in the King County real estate market.

Data Understanding.

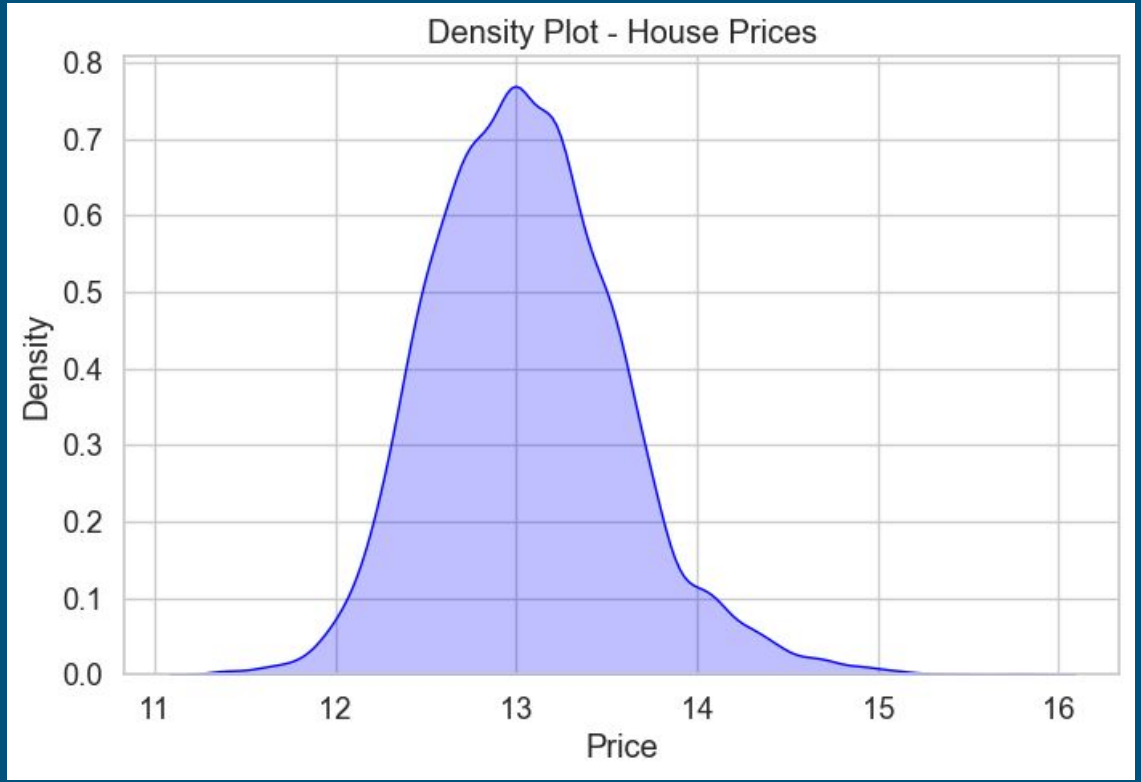
The dataset used for our project contains information on house sales in King County. It includes columns such as `id`, `date`, `price`, `bedrooms`, `bathrooms`, `sqft_living`, `sqft_lot`, `floors`, `waterfront`, `view`, `condition`, `grade`, `sqft_above`, `sqft_basement`, `yr_built`, `yr_renovated`, `zipcode`, `lat`, and `long`. These columns provide valuable insights into various aspects of the properties, including their physical attributes, location-related features, and sale prices. By analyzing these variables, we can gain a comprehensive understanding of the housing market in King County and identify the factors that significantly impact house sale prices.

Data Preparation.

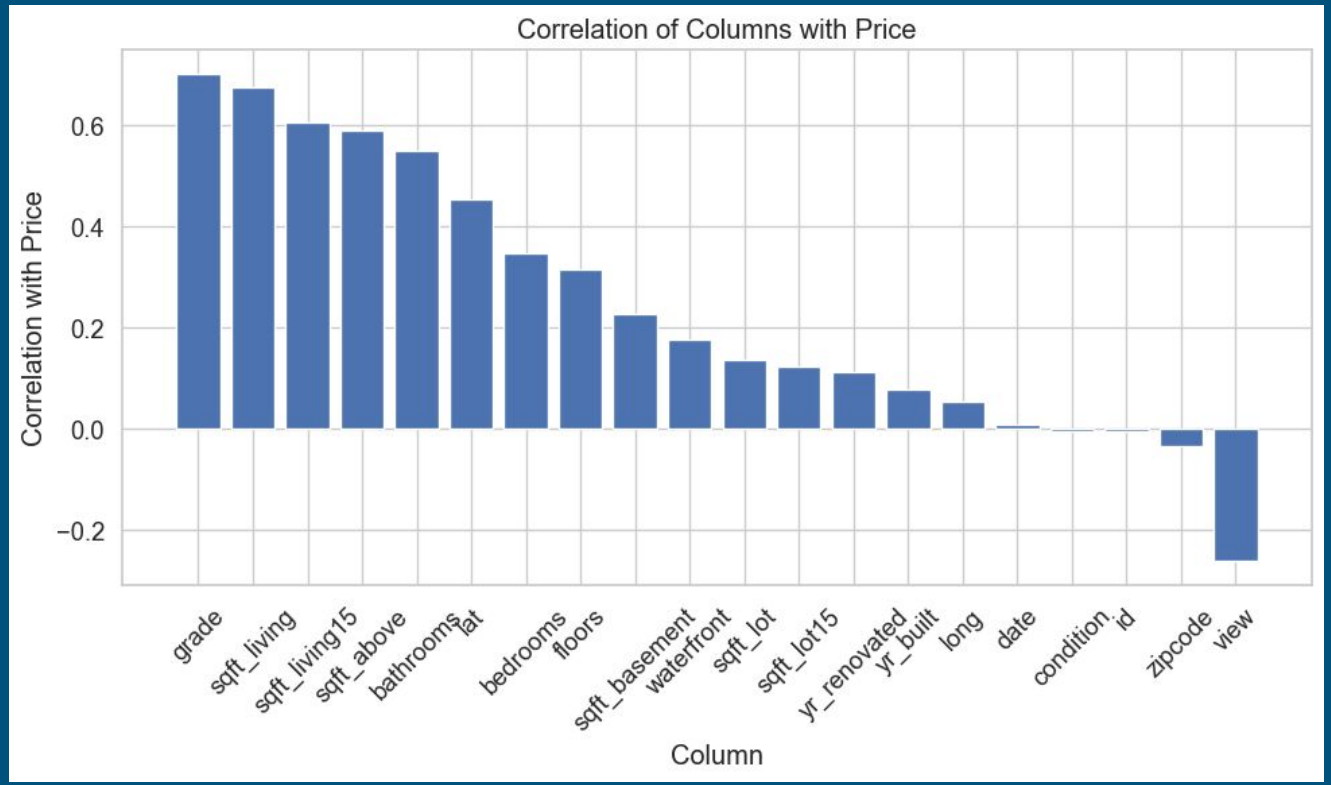
1. We handled missing data by dropping rows with NaN values.
2. We handled categorical data by use of label encoding.
3. Found columns that were skewed and applied log transformation to them.
4. We found outliers in our dataset and removed the relevant rows.
5. We performed data standardization for dimension reduction

Graph showing Price Density before fixing data skewness

Graph showing price Density after fixing data skewness.



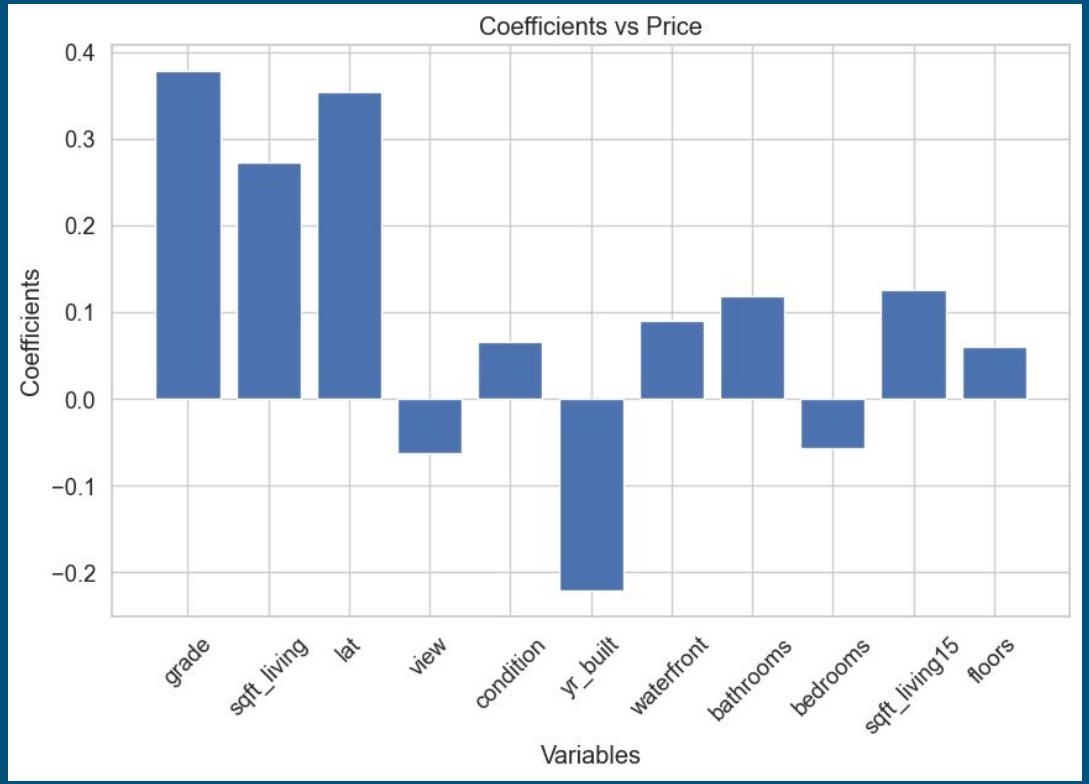
Correlation of all columns in data to Price



Analysis of column correlation to price

The correlation analysis reveals that certain factors have a stronger influence on house prices than others. Variables such as sqft_living, grade, and sqft_above show the strongest positive correlations with price, indicating that larger and higher-quality houses tend to command higher prices. The number of bathrooms and the latitude of the location also display relatively strong positive correlations. On the other hand, variables like view and waterfront have weaker correlations, suggesting that they have a smaller impact on prices. Overall, these findings highlight the importance of size, quality, and location in determining house prices in the dataset.

Coefficients of all chosen independent variables



Analysis of coefficients for the chosen independent variables.

The coefficients in the regression model represent the estimated effects of the independent variables on the dependent variable (price). The analysis of the coefficients reveals the following associations with the price: a higher grade of a property, larger square footage of the living area, properties located at higher latitudes, better condition, being waterfront, more bathrooms, nearby properties with larger living areas, and more floors are positively associated with higher prices. On the other hand, a better view, older properties, and more bedrooms are associated with lower prices. These coefficients provide valuable insights into the factors that influence house prices in the dataset.

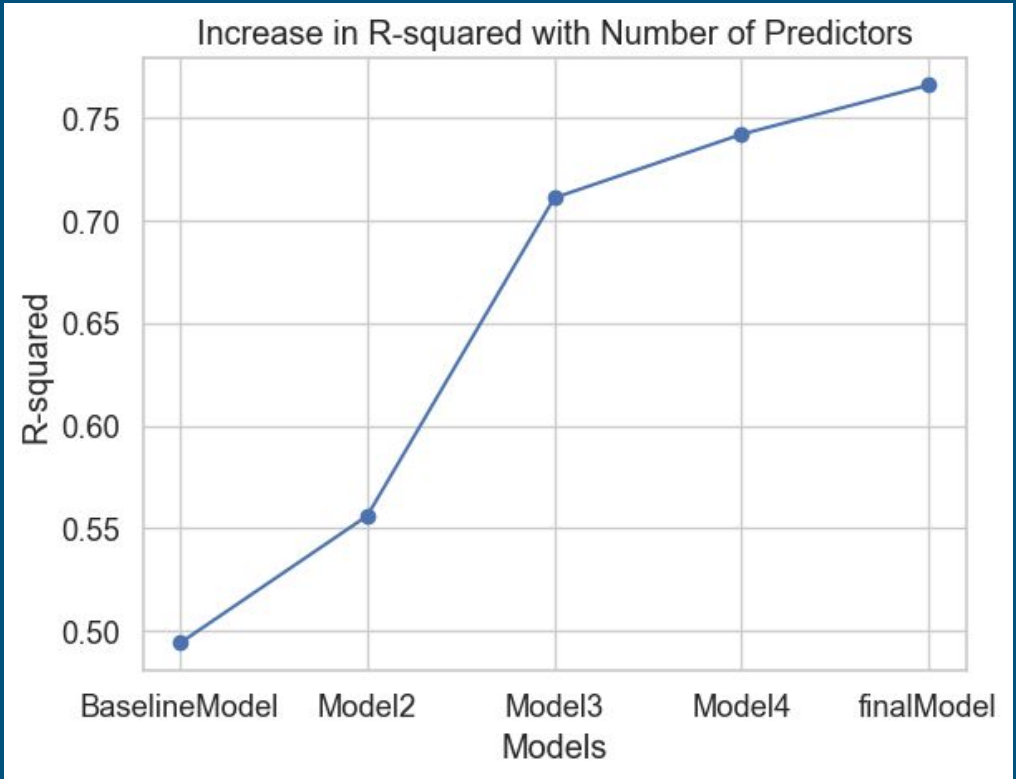
Modelling.

Our training strategy was to create a baseline model and from there we try to improve the r squared by adding more independent variables

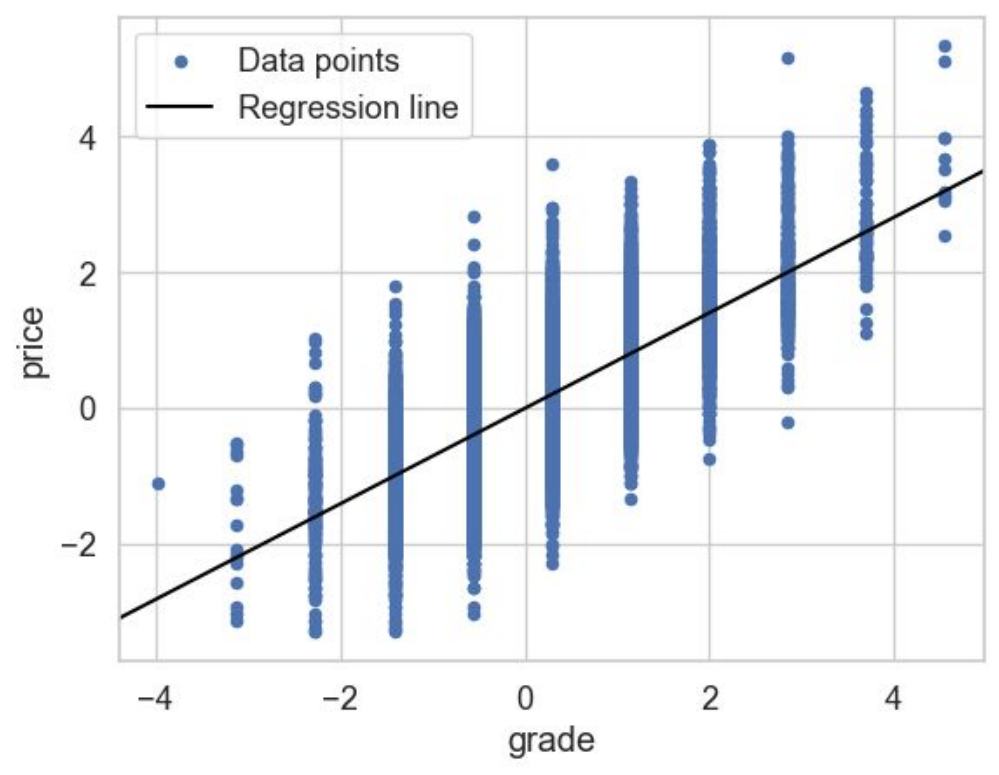
We created 5 models in total:

1. Baseline Model.(one independent variable)
2. Model 2 (two independent variables)
3. Model3 (4 independent variables)
4. Model4 (6 independent variables)
5. Final model (11 independent variables)

Graph showing each Model's Performance



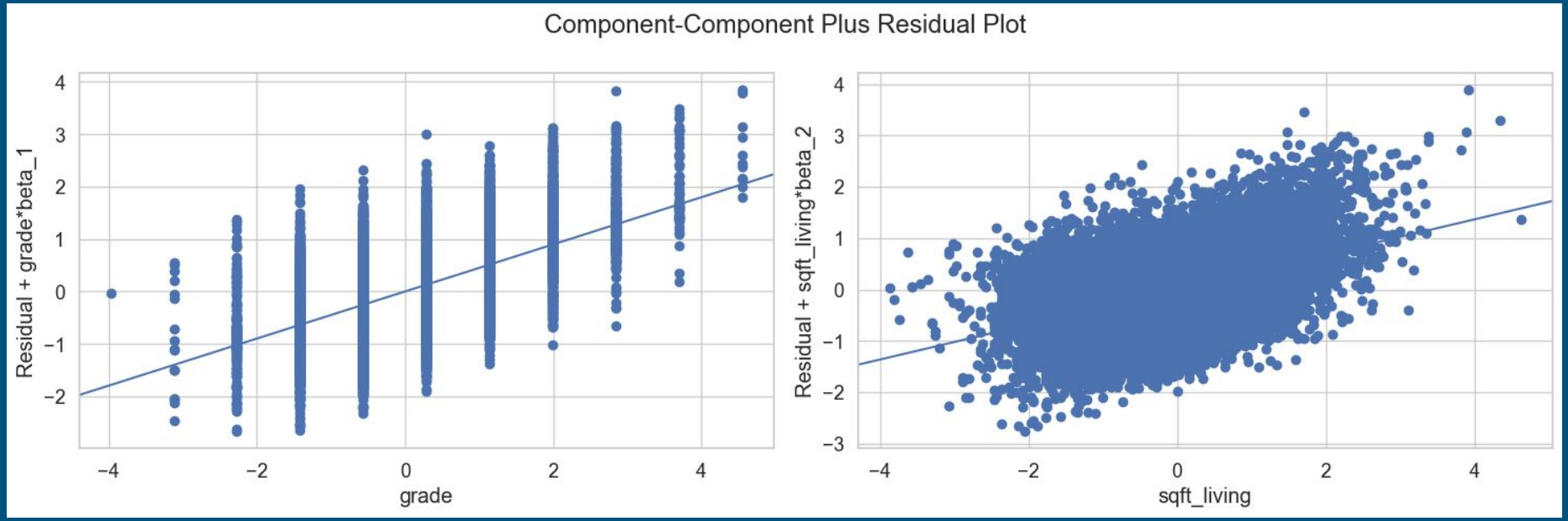
Baseline model



Summary of Baseline Model analysis

The baseline model, with only the grade variable, shows a moderate level of goodness-of-fit, with an R-squared value of 0.494. This indicates that approximately 49.4% of the variation in the price can be explained by the grade. The F-statistic is very high, indicating that the model is statistically significant. The coefficient for the grade variable is 0.7032, suggesting that for each unit increase in grade, there is a corresponding increase in the price by 0.7032 units. The coefficient is statistically significant, as indicated by the low p-value and high t-statistic. However, further analysis is needed to assess the model's assumptions and consider the inclusion of additional variables for better performance.

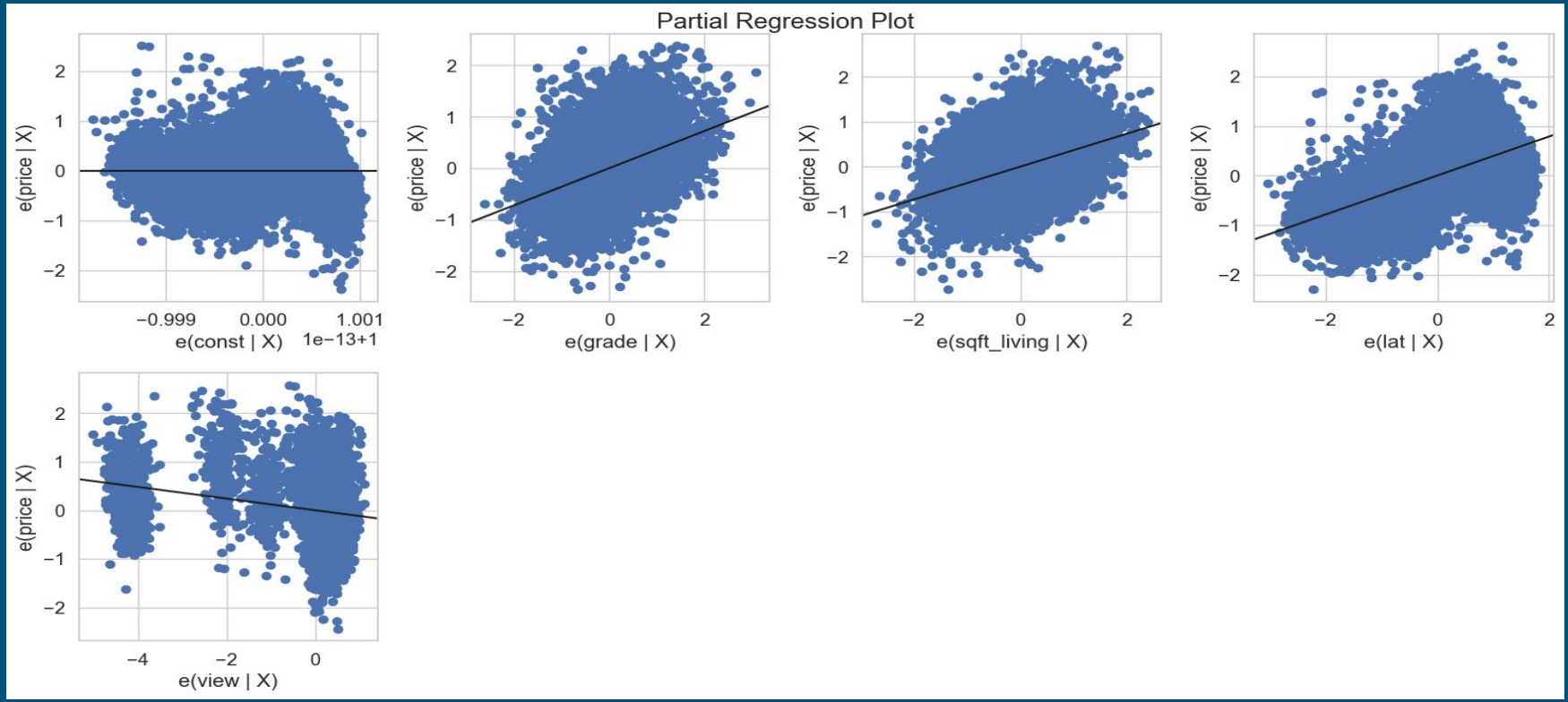
Model 2


Component-Component Plus Residual Plot
Residual + grade*beta_1
grade
Residual + sqft_living*beta_2
sqft_living

Summary of Model 2 analysis

Model 2, which includes the grade and sqft_living variables, shows a moderate level of goodness-of-fit, with an R-squared value of 0.546. This indicates that approximately 54.6% of the variation in the price can be explained by these two variables. The adjusted R-squared value is also 0.546, suggesting that the model adequately accounts for the complexity of the predictors. The F-statistic is high, indicating that the model is statistically significant. The coefficients for both grade and sqft_living are statistically significant and have positive relationships with the price. Further analysis is required to assess the model's assumptions and explore the inclusion of additional variables for potential improvement.

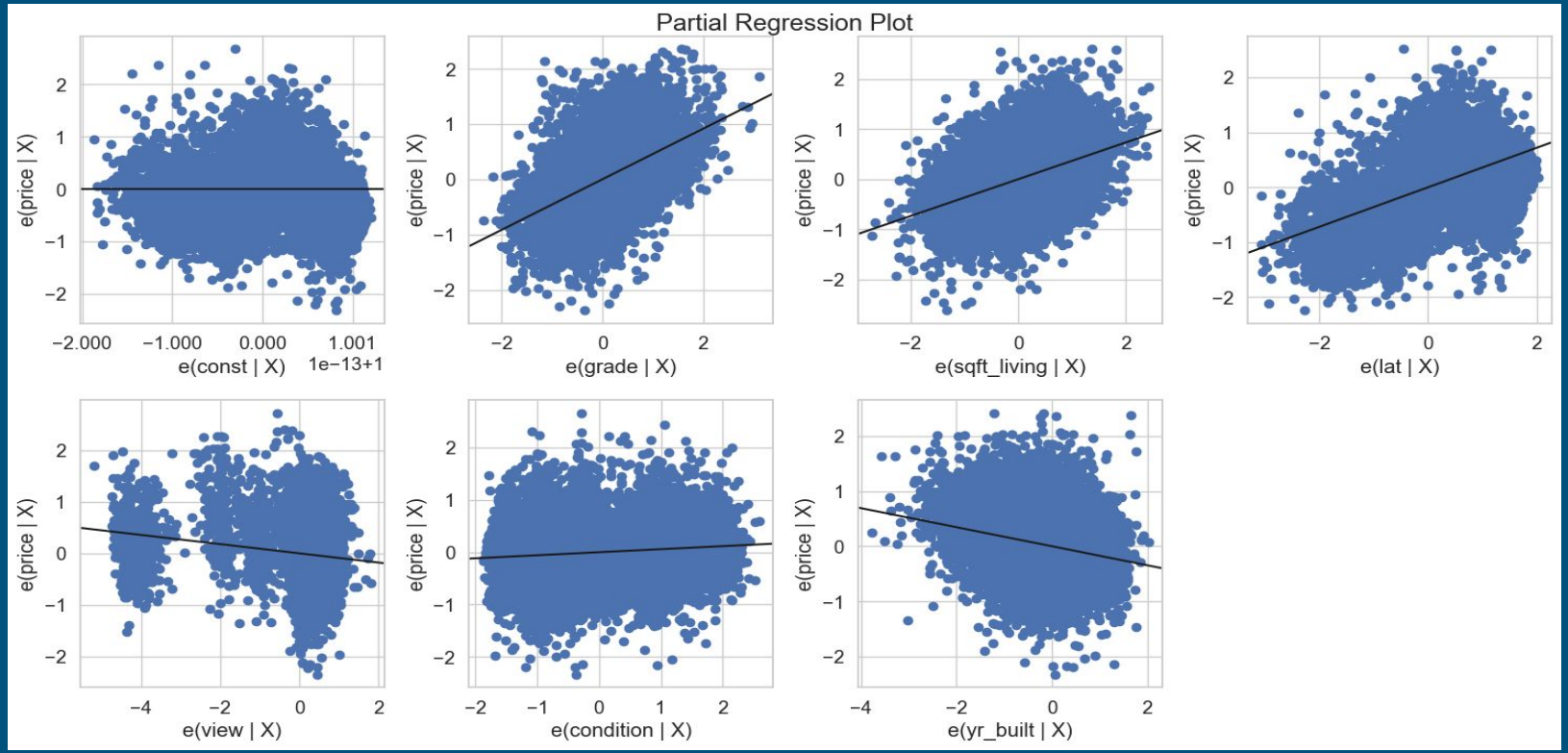
Model 3


Partial Regression Plot
e(price | X)
e(const | X)
−0.999
0.000
1.001
1e−13+1
e(grade | X)
e(sqft_living | X)
e(lat | X)
e(view | X)

Summary of analysis of model 3

Model 3, which includes the grade, sqft_living, lat, and view variables, shows a relatively high level of goodness-of-fit, with an R-squared value of 0.711. This indicates that approximately 71.1% of the variation in the price can be explained by these variables. The adjusted R-squared value is also 0.711, suggesting that the model adequately accounts for the complexity of the predictors. The F-statistic is high, indicating that the model is statistically significant. The coefficients for all the variables are statistically significant and have meaningful interpretations in relation to the price. Further analysis is required to assess the model's assumptions and explore the inclusion of additional variables for potential improvement.

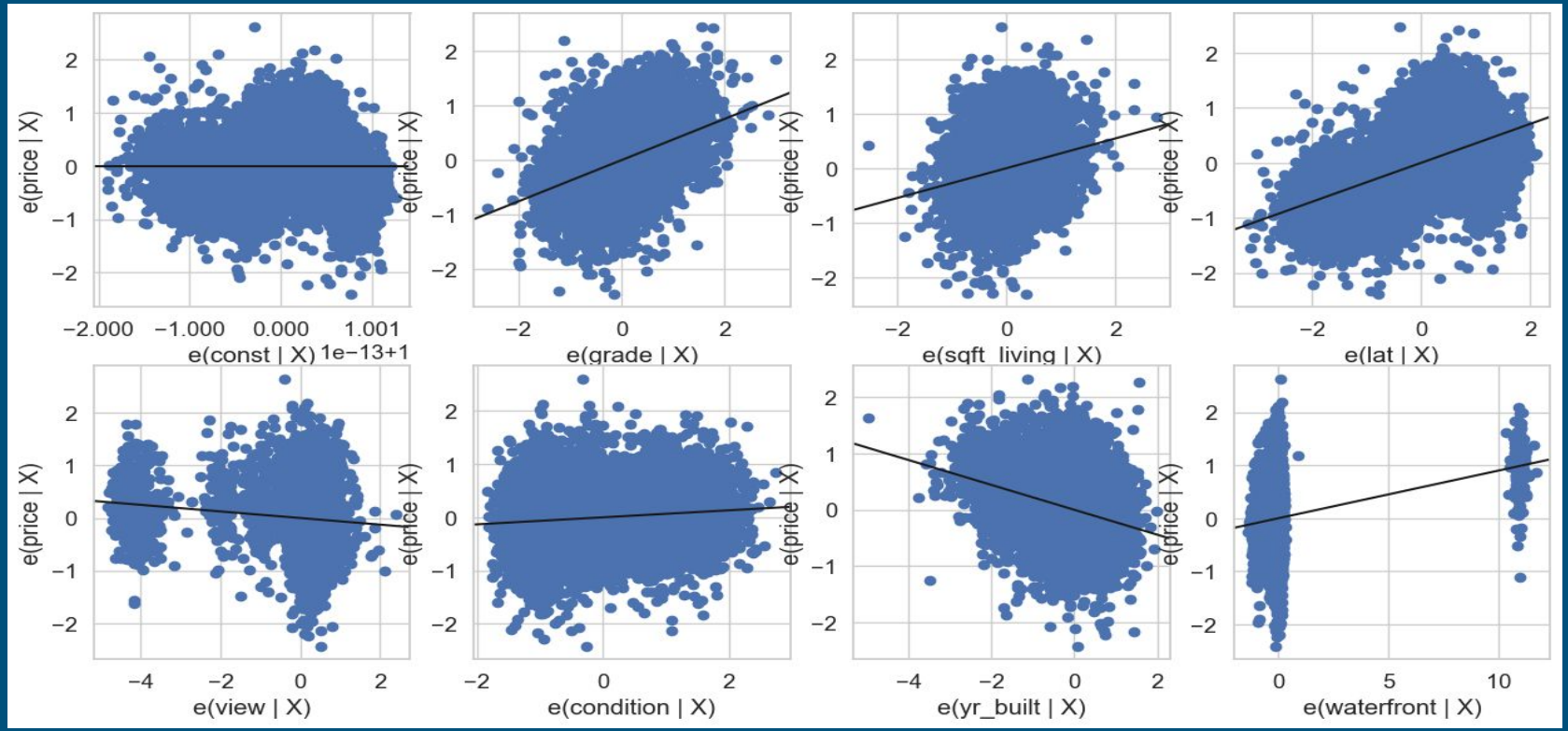
Model 4


Partial Regression Plot
e(price | X)
e(const | X)
1e−13+1
e(grade | X)
e(sqft_living | X)
e(lat | X)
e(view | X)
e(condition | X)
e(yr_built | X)

Summary of Model 4 analysis

Model 4 includes the grade, sqft_living, lat, view, condition, and yr_built variables. The analysis shows a relatively high level of goodness-of-fit, with an R-squared value of 0.742. This means that approximately 74.2% of the variation in the price can be explained by these variables. The adjusted R-squared value is also 0.742, indicating that the model adequately accounts for the complexity of the predictors. The F-statistic is high, indicating that the model is statistically significant. All the coefficients are statistically significant and have meaningful interpretations in relation to the price. Further analysis is required to assess the model's assumptions and explore the inclusion of additional variables for potential improvement.

Final Model


e(price | X)
e(const | X)
1e−13+1
e(grade | X)
e(sqft_living | X)
e(lat | X)
e(view | X)
e(condition | X)
e(yr_built | X)
e(waterfront | X)

Summary for final model analysis

The final model analysis reveals a strong fit with an R-squared value of 0.766, indicating that 76.6% of the price variation can be explained by the included independent variables. The F-statistic confirms the overall model's significance, and all coefficients are statistically significant with p-values below 0.05. While the normality assumption is slightly violated based on the Omnibus and Jarque-Bera tests, the residuals exhibit a relatively symmetrical distribution and moderate peakedness. The condition number suggests a moderate level of multicollinearity, and the Durbin-Watson statistic indicates no significant autocorrelation in the residuals. Overall, the final model provides a solid foundation for predicting house prices based on the selected independent variables.

Project Summary

The multiple linear regression model achieved a high R-squared value of 0.766, indicating that approximately 76.6% of the variability in housing prices can be explained by the selected predictors. Factors such as grade, sqft_living, lat, waterfront, and bathrooms have positive coefficients and positively impact the price, while variables like view, condition, yr_built, bedrooms, sqft_living15, and floors have negative coefficients and negatively affect the price. Improving the property's grade and considering factors like living area and location can increase the price, while buyers can find better value by considering factors with a negative association. However, it's important to be mindful of the model's assumptions and limitations when applying these findings in real-world scenarios.

Recommendations to client 1 : Seller.

1. Improve the overall grade of the property.
2. Enhance the living space by increasing the square footage.
3. Highlight the desirable aspects of the property's location.
4. Maximize the property's view if applicable.
5. Ensure the property is in good condition and well-maintained.
6. Consider renovations to update older properties.
7. Emphasize waterfront features if the property has them.
8. Enhance bathrooms, bedrooms, and other living spaces.

Recommendations to client 2 : Buyer

- Focus on properties with a suitable grade and size.
- Consider desirable locations with amenities.
- Prioritize properties in good condition.
- Evaluate the number of bedrooms, bathrooms, and living space layout.
- Seek advice from a local real estate professional.
- Consider the age of the property and potential for renovations.
- Explore properties near waterfront areas if desired.
- Make an informed decision based on budget and personal preferences.