Data-Driven Renovation Advice for Homeowners in King County

OVERVIEW

Business Problem

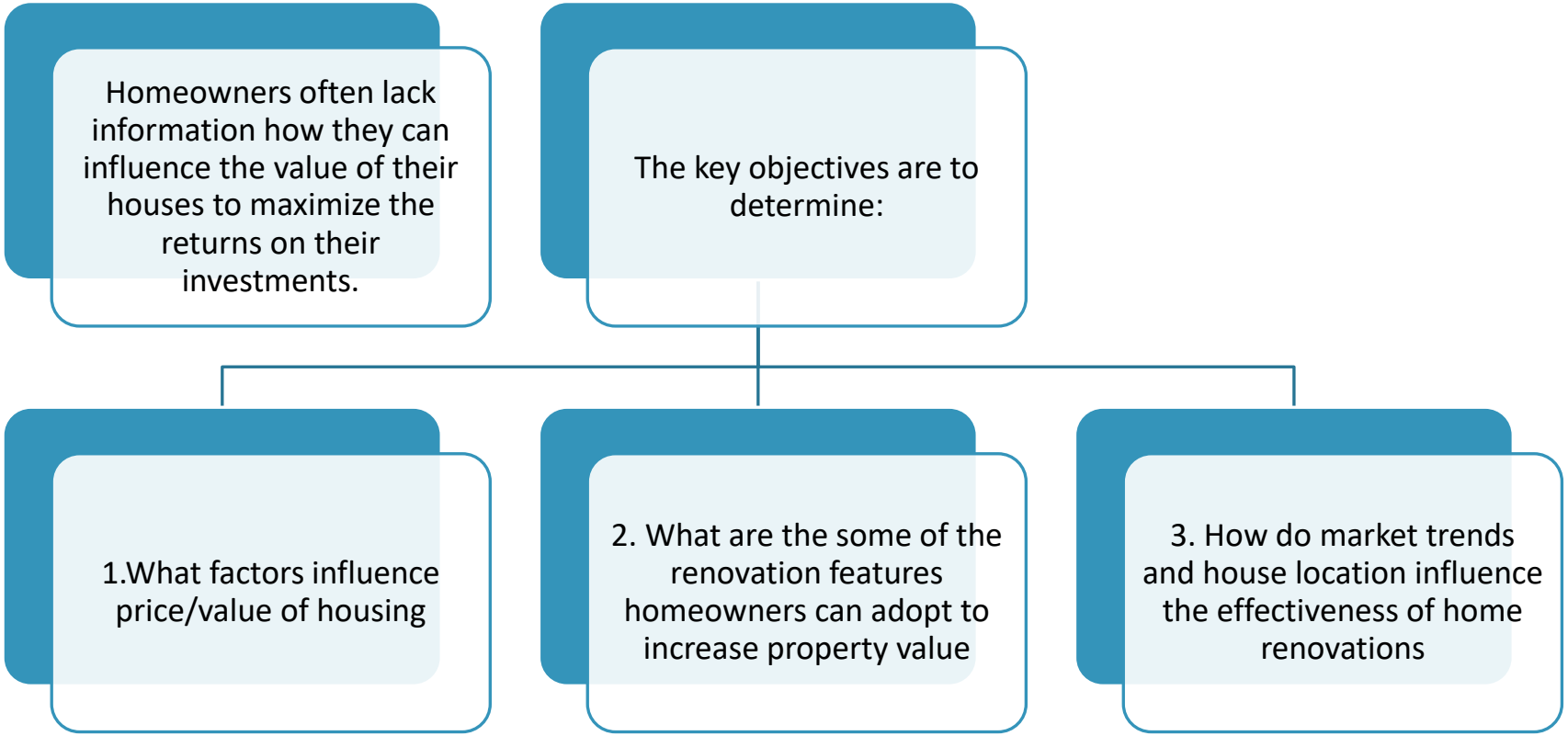
Data Understanding

Modelling

Observations

Recommendations and Next Steps

BUSINESS PROBLEM



DATA UNDERSTANDING

This project uses the King County House Sales dataset.

The description of the features under review include :

	Column Names and Descriptions for King County Data Set
1	`id` - Unique identifier for a house
2	`date` - Date house was sold
3	`price` - Sale price (prediction target)
4	`bedrooms` - Number of bedrooms
5	`bathrooms` - Number of bathrooms
6	`sqft_living` - Square footage of living space in the home
7	`sqft_lot` - Square footage of the lot
8	`floors` - Number of floors (levels) in house
9	`waterfront` - Whether the house is near a waterfront e.g Includes Duwamish, Elliott Bay, Puget Sound, Lake Union, Ship Canal, Lake Washington, Lake Sammamish, other lake, and river/slough waterfronts
11	`view` - Quality of view from house e.g Includes views of Mt. Rainier, Olympics, Cascades, Territorial, Seattle Skyline, Puget Sound, Lake Washington, Lake Sammamish, small lake / river / creek, and other
13	`condition` - How good the overall condition of the house is. Related to maintenance of house.
15	`grade` - Overall grade of the house. Related to the construction and design of the house.
17	`sqft_above` - Square footage of house apart from basement
18	`sqft_basement` - Square footage of the basement
19	`yr_built` - Year when house was built
20	`yr_renovated` - Year when house was renovated
21	`zipcode` - ZIP Code used by the United States Postal Service
22	`lat` - Latitude coordinate
23	`long` - Longitude coordinate
24	`sqft_living15` - The square footage of interior housing living space for the nearest 15 neighbors
25	`sqft_lot15` - The square footage of the land lots of the nearest 15 neighbors

Methods

Data Preparation

Preview of the data set on hosing in Northwestern County.

Review of the data to get a sense of its shape and structure.

Check on the data types of each column and handling any inconsistencies.

Check for missing values in each column.

FUTURE SELECTION

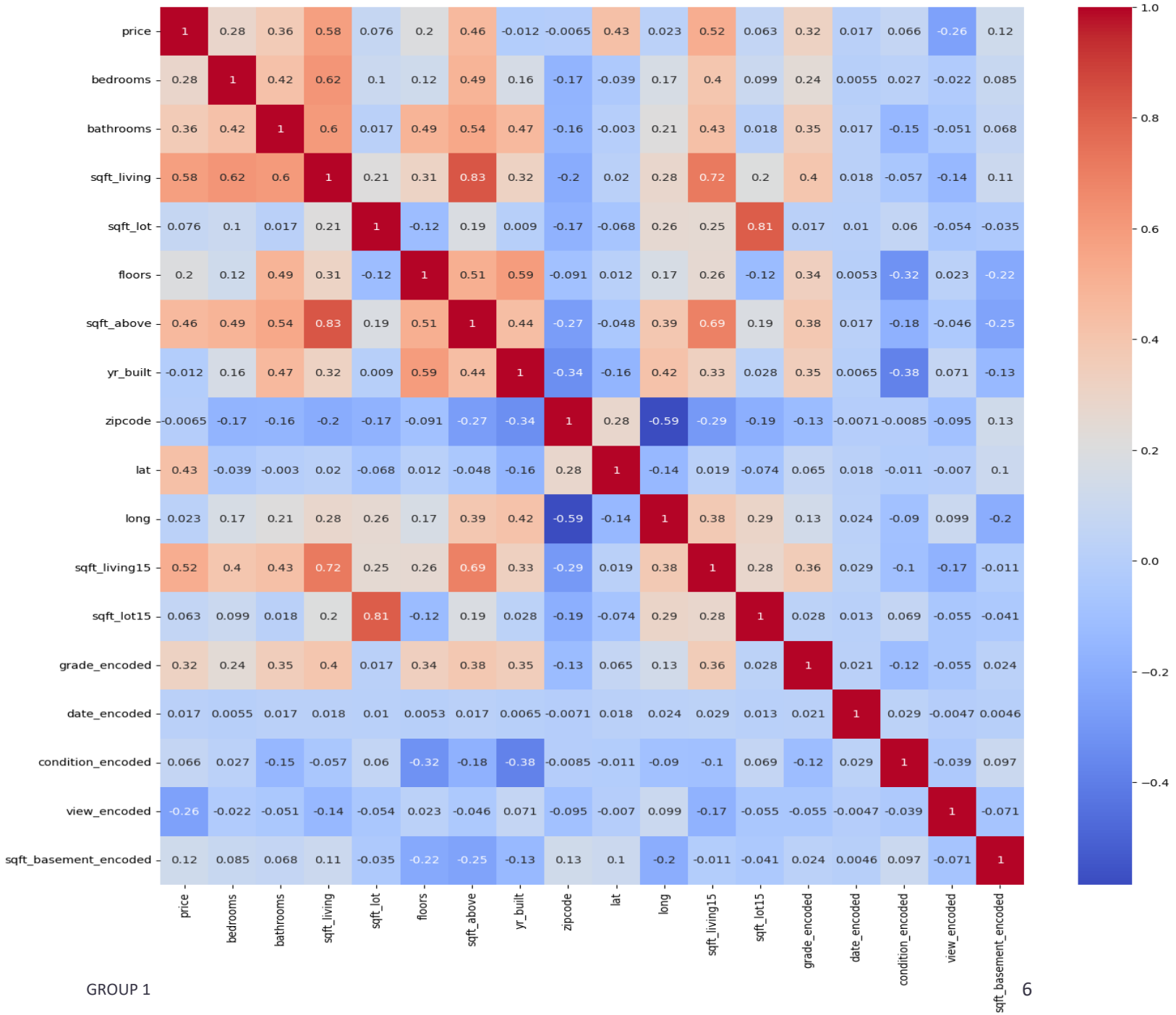
- Based on the correlation coefficients between various features and the target variable (price), we can draw the following conclusions:

1. Strong Positive Correlation:

- sqft_living (0.583466): The living area square footage has a strong positive correlation with the price (as target variable.)
- sqft_living15 (0.522030): Similarly, the square footage of interior living space for the nearest 15 neighbors has a strong positive correlation with the price.
- sqft_above (0.464570): The square footage of house apart from the basement has a strong positive correlation with the price.

2. Moderate positive correlation

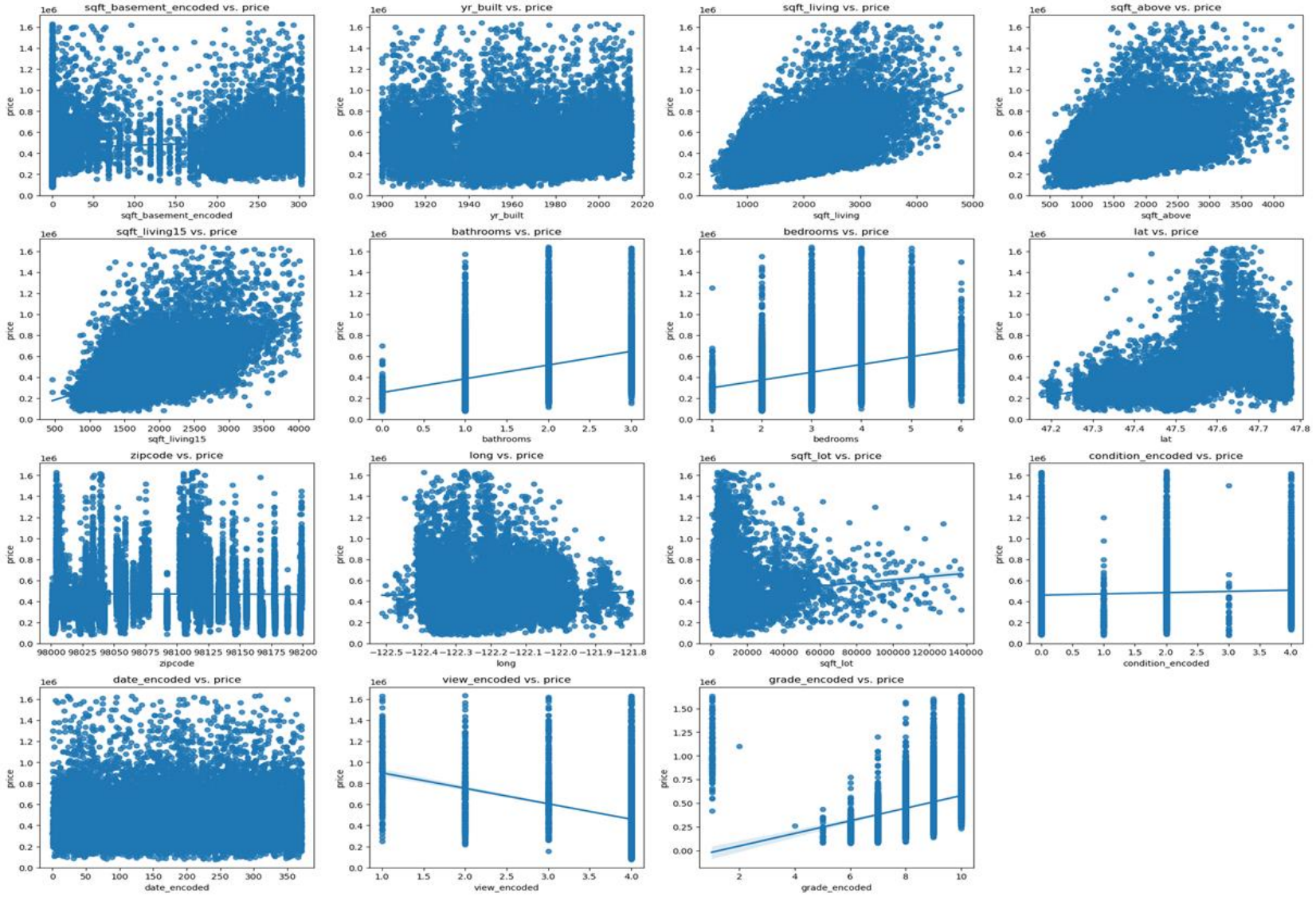
- Bathrooms (0.362375): The number of bathrooms shows a moderate positive correlation with the price.
- Grade_encoded (0.318128): The encoded grade of the property also exhibits a moderate positive correlation with the price.



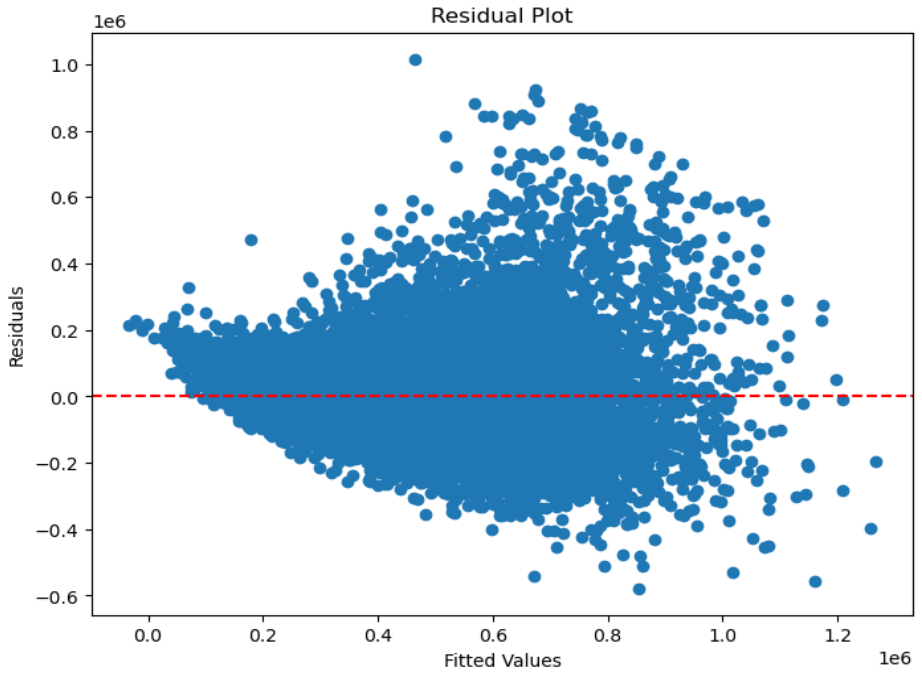
ASSUMPTIONS OF LINEAR REGRESSION

1. Linearity

- There is a positive linear correlation relationship between price and sqft_living,sqft_above,sqft_living15, and bathrooms.
- There is a negative linear correlation relationship between price and grade_encoded, view, encoded and zipcode.
- From the model, lat proved to have a stronger linearity with price, which determines the location homeowners could venture into.
- There is a weak linear correlation relationship between price and independent variables- long, condition_encoded, yr_built, sqft_basement encoded, sqft_lot15, sqft_lot, yr_renovated, floors, waterfront_encoded and bedrooms.

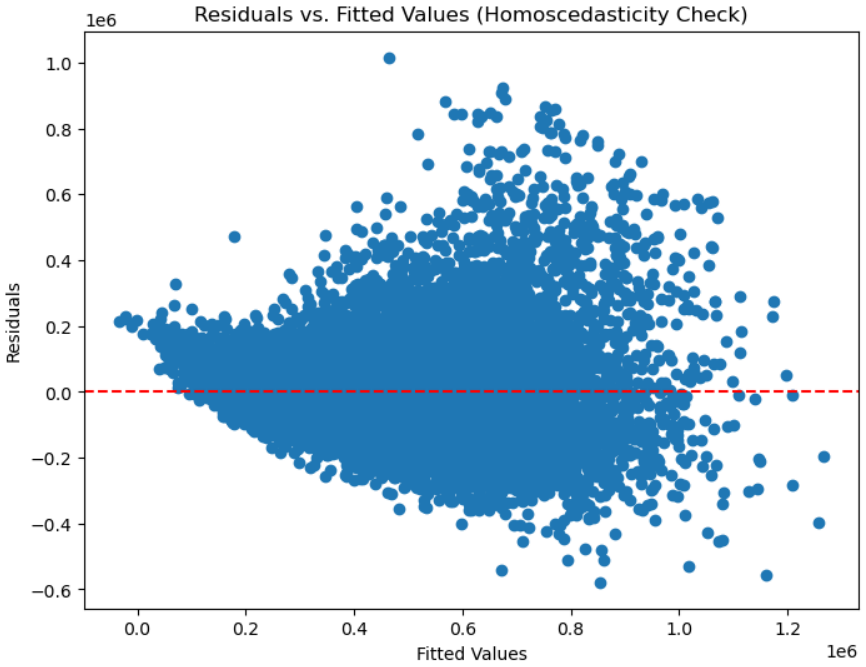


2. Independence



The above test statistic of approximately 1.99 indicates that there is very little evidence of autocorrelation in the residuals. This suggests that the independence assumption in linear regression may be reasonable for the model. Therefore Providing evidence that the independence assumption in linear regression is not violated, indicating that the model's residuals exhibit no significant autocorrelation pattern.

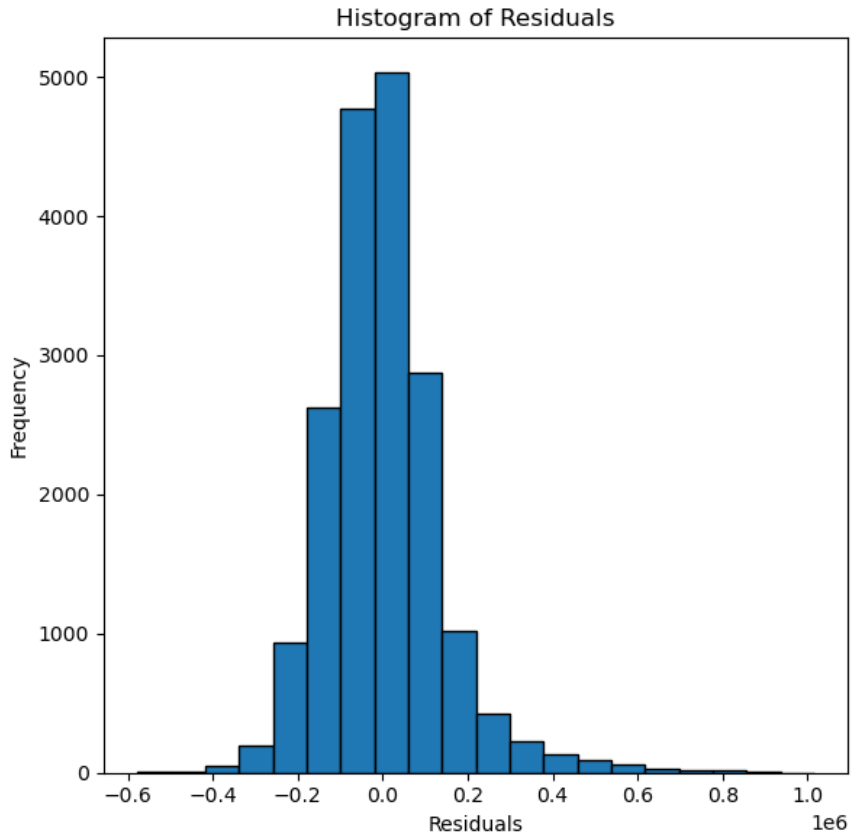
3. Homoscedasticity

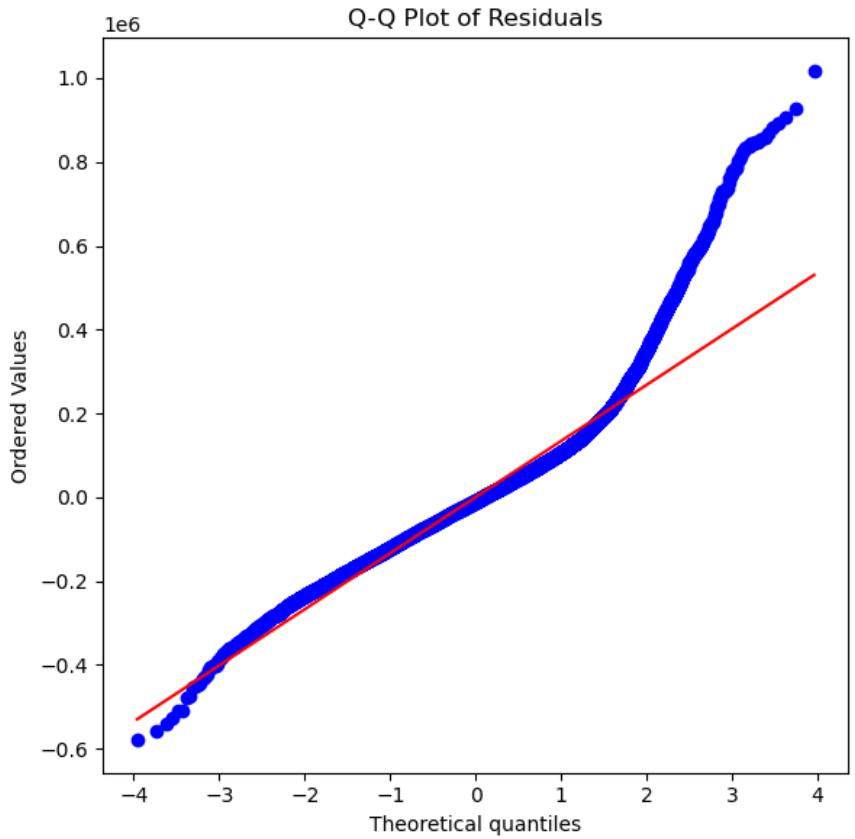


The observed pattern is a Cone-shaped or the spread of residuals increases or decreases systematically as the predicted values change. This indicates heteroscedasticity, violating the assumption of constant variance homoscedasticity

4. Normality

Histogram has a normal Distribution and has a Right-tail. The Q-Qplot is also not heavily skewed.





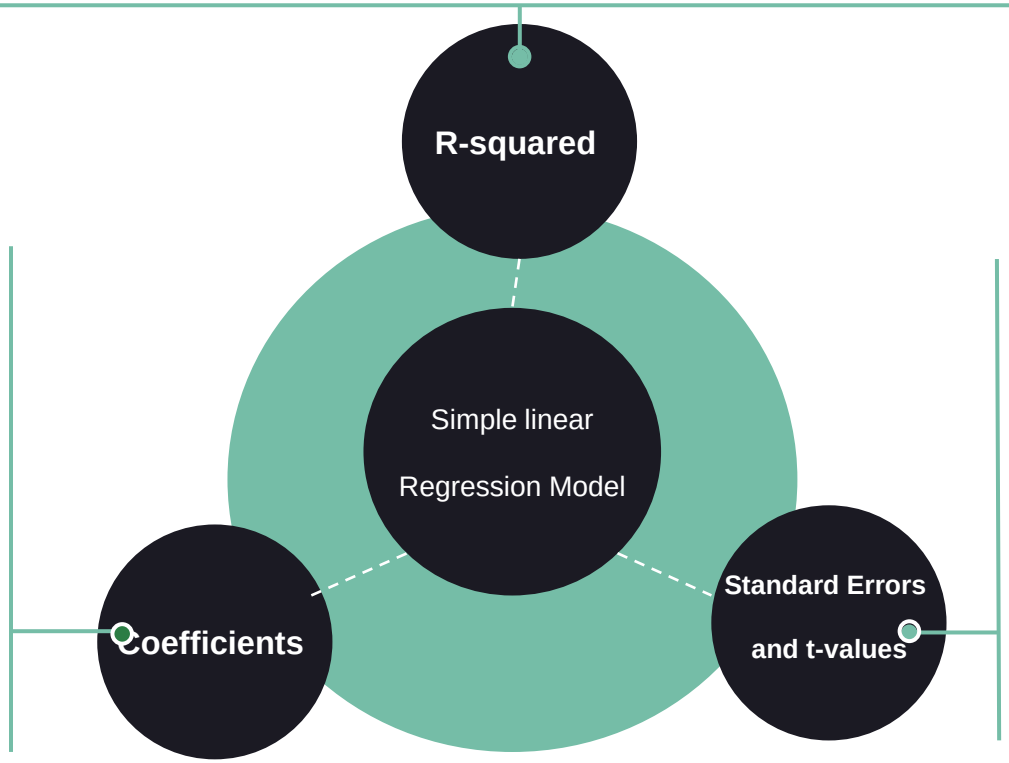
MODELLING

1. Base model: Simple Linear

☑ **In this model, the R-squared value is 0.340, indicating that approximately 34% of the variability in house prices can be explained by the square footage of living space.**

☑ **The coefficient for the intercept term is approximately 115,800, indicating the estimated average house price when the square footage of living space is zero (which is not practically meaningful)**

☑ **The coefficient for the sqft_living variable is approximately 186.45, indicating that, on average, each additional square foot of living space is associated with an increase in house price.**



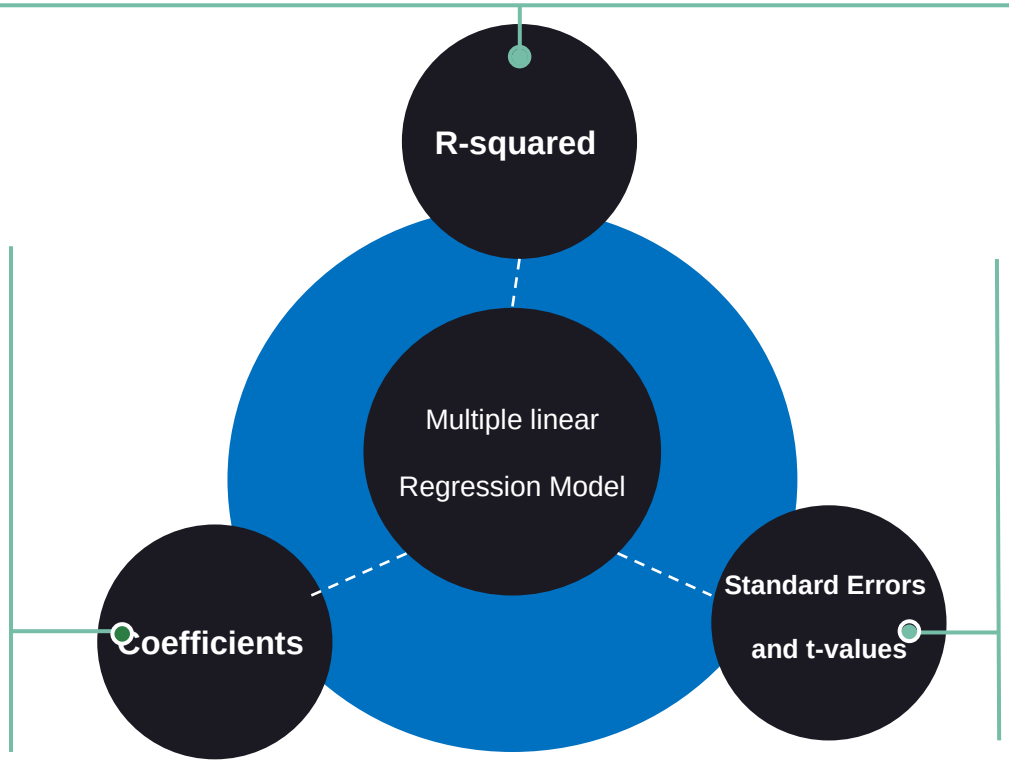
☑ **Both the intercept and sqft_living coefficients have very low p-values ($p < 0.001$), indicating that they are statistically significant predictors of house price.**

2. Multiple Regression

☑ **In this model, the R-squared value is 0.618, indicating that approximately 61.8% of the variability in house prices can be explained by the independent variables included in the model.**

☑ The coefficients represent the estimated effect of each independent variable on the dependent variable, holding other variables constant.

☑ For example, the coefficient for sqft_living is approximately 113.28, indicating that a one-unit increase in square footage of living space is associated with an increase in house price of approximately $113.28, holding other variables constant.



☑ Standard errors estimate the variability of the coefficient estimates.

☑ The t-values indicate the significance of each coefficient. The p-values associated with each coefficient test the null hypothesis that the coefficient is equal to zero. In this case, most coefficients have very low p-values ($p < 0.001$), indicating that they are statistically significant predictors of house price.

3. Random Forest Regression Model

- **Mean Squared Error (MSE)**

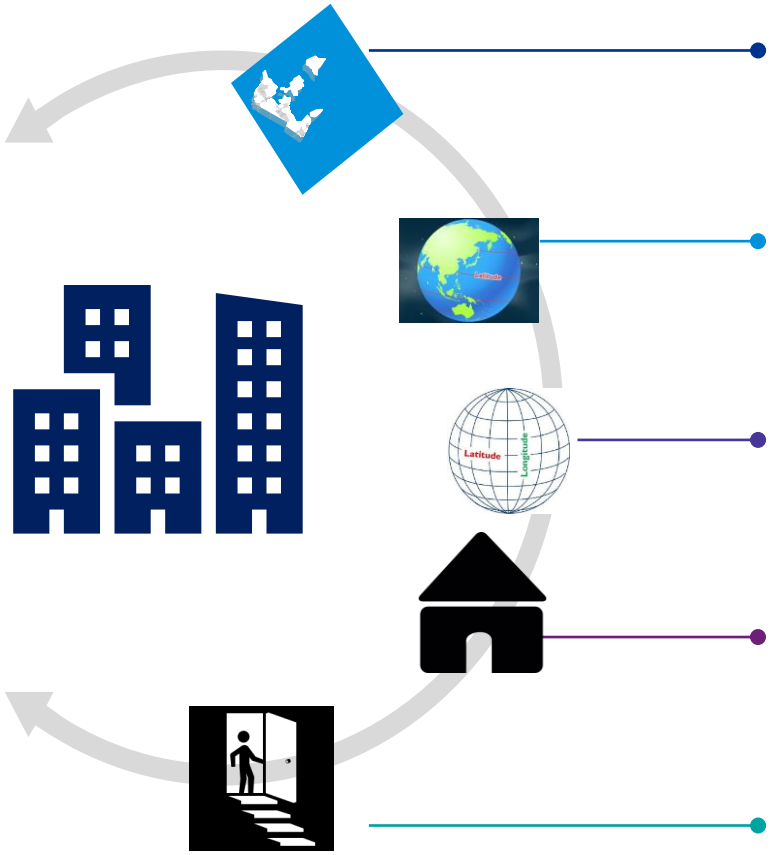
The MSE is a measure of the average squared difference between the actual and predicted values. In this case, the MSE is approximately 7.89 billion. Lower MSE values indicate better fit, meaning that the model's predictions are closer to the actual values on average.

- **R-squared**

The R-squared value is approximately 0.847, which means that around 84.7% of the variance in the dependent variable (target) is explained by the independent variables (features) in the model. This is a relatively good R-squared value, indicating that the model fits the data well.

- **Overall, Random Forest Regression model seems to perform reasonably well based on these evaluation metrics. These performance metrics suggest that the Random Forest Regression model performs relatively well in predicting house prices based on the given features.**

Main Observations



sqft_living:
This feature has the highest importance with a value of approximately 0.314. It suggests that the square footage of living space is the most influential feature in predicting house prices.

lat:
The latitude of the location comes next in importance, with a value of approximately 0.398. This indicates that the geographical location, represented by latitude, plays a significant role in determining house prices.

long:
The longitude of the location follows, with a value of approximately 0.062. Longitude is also an important geographical feature in predicting house prices.

sqft_living15:
This feature represents the average square footage of interior housing living space for the nearest 15 neighbors. Its importance is approximately 0.054.

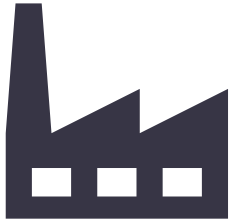
sqft_basement_encoded:
The encoded indicator of whether the house has a basement has an importance value of approximately 0.006. condition_encoded, bedrooms, bathrooms, view_encoded, floors: These features have relatively lower importance values ranging from approximately 0.001 to 0.006.

RECOMMENDATIONS

Recommendations

It is important to acknowledge a significant limitation of this analysis: the dataset used is specific to King County, a single county within a major metropolitan area. Consequently, the insights derived may not be directly applicable to other cities or counties.

Location of the house



- It is essential to recognize that the size of the property emerges as the most significant predictor of its price. Homeowners may consider increasing the square footage of living space as it is comparatively competitive and a presumed commanding source of revenue.

Region specific

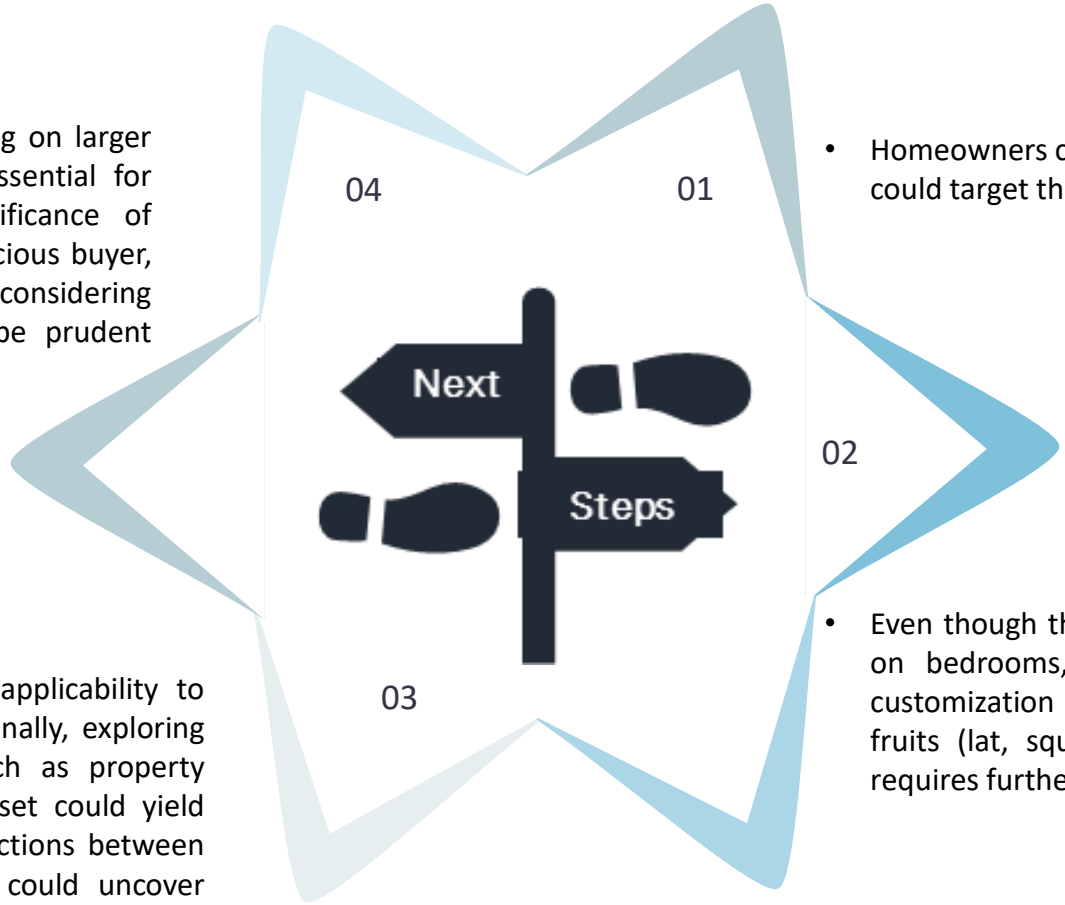
- Homeowners can choose to renovate houses near the tropical climate areas/equatorial geographic regions as it would attract more income.

Focus on High-Impact Renovations

- Prioritize renovations that have the highest impact on property values, such as increasing square footage of living space (sqft_living). This feature has been identified as the most influential factor in predicting house prices.

Next Steps

- For developers operating in King County, focusing on larger builds while ensuring quality construction is essential for maximizing sale prices. Recognizing the significance of location cannot be overstated. As a budget-conscious buyer, exploring areas with potential for growth and considering properties that offer value for money could be prudent strategies to pursue.



- Homeowners could renovate houses in these regions as they could target the tourism industry which may have better ROI.

- It would be beneficial to evaluate the model's applicability to various locations across the United States. Additionally, exploring alternative predictions beyond house prices, such as property valuation or market trends, using the same dataset could yield valuable insights. Furthermore, investigating interactions between variables and exploring polynomial relationships could uncover additional nuances and improve the accuracy of the model.

- Even though there is a lower output in property value based on bedrooms, bathrooms, view and floors, would their customization serve as additive value for the low-hanging fruits (lat, square footage of living space and long). This requires further exploration.

Thank You!