Predicting King County Housing Prices

Group Members

1. Julius Charles

2. Brenda Kinoti

3. Cleophas Opati

4. Winnie Onduru

5. John Karanja

6. Joy Wangari

7. Jonathan Okwaro

Business Understanding

OBJECTIVE:

Develop a model for precise house price estimates in King County. Provide reliable information for buyers and sellers.

APPROACH:

Build a model to deduce main factors influencing house prices. Utilize accurate and representative real estate data. Consider historical sales, current listings, property size, and other relevant features.
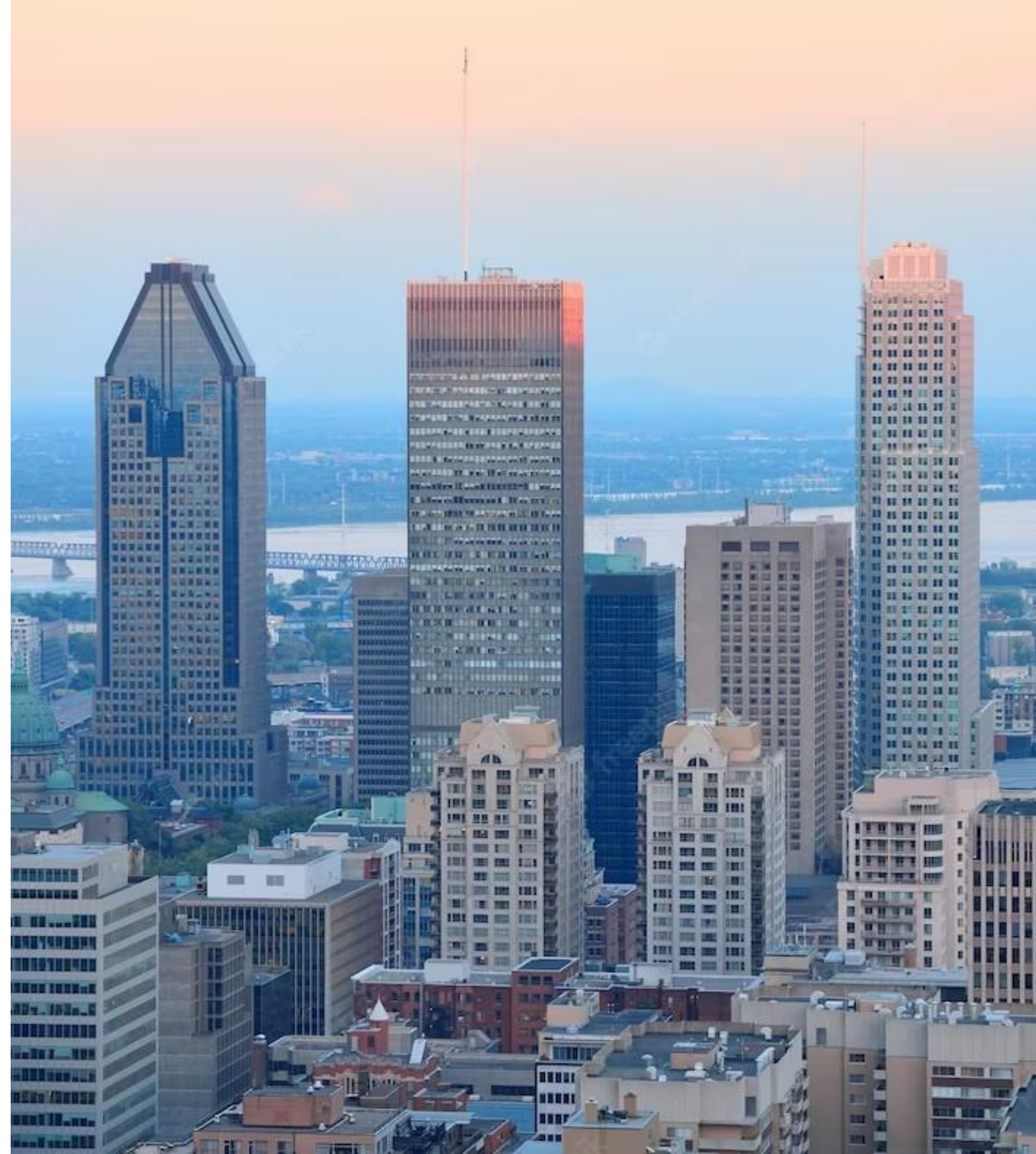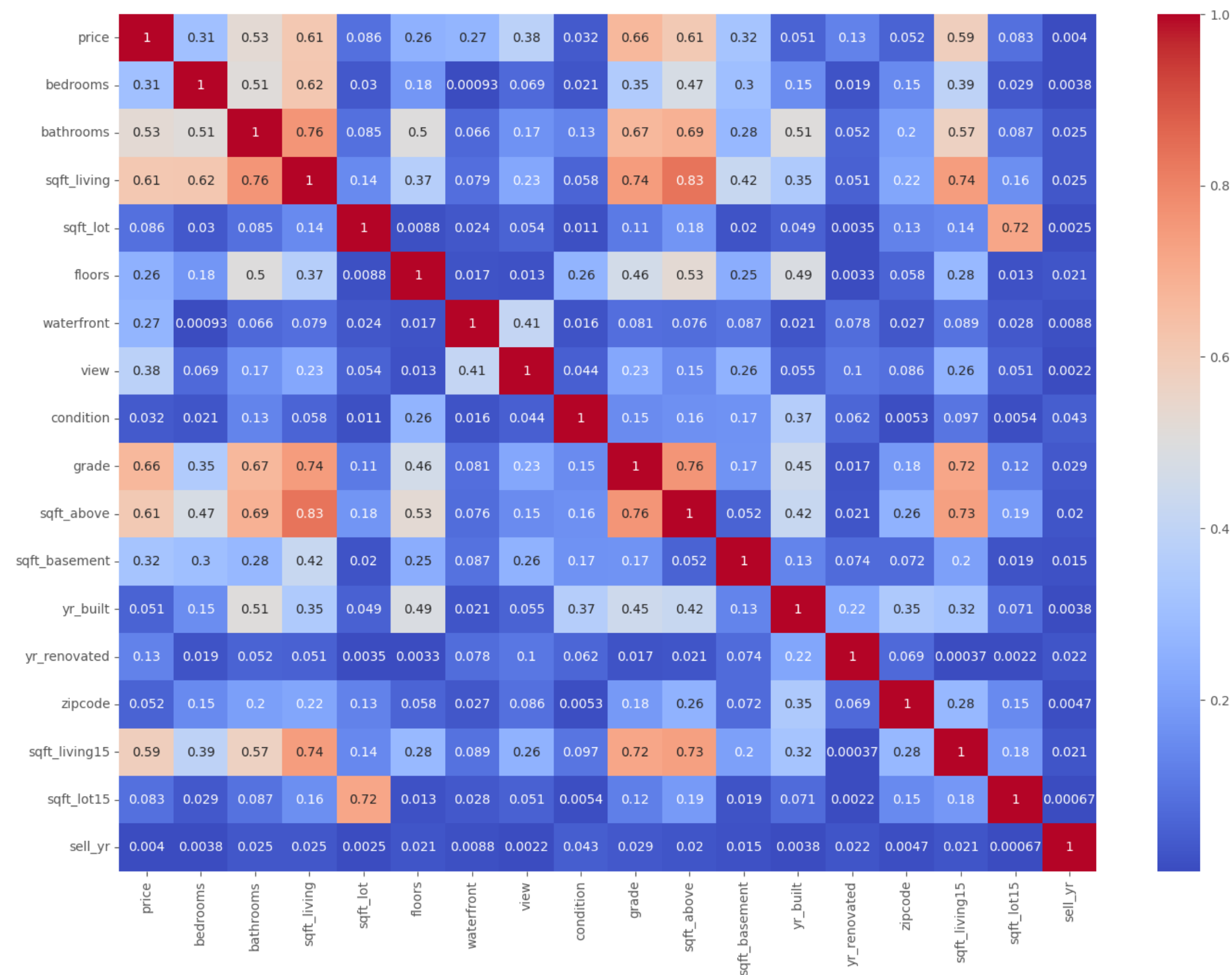
Data Sources

For this ANALYSIS, we will be using the King County House Sales Dataset, which contains Information about house sales in a northwestern county.
This data was collected between 2014 to 2015.

Correlation Matrix of King County Housing data

# Question 1: What Features Have The Highest Correlation To The Home Price

In general, a correlation value above 0.7 is considered high. Although the dataset does not contain any correlations with the price above 0.7, there are several features that exhibit moderately strong correlations. The variables sqft_living, grade, sqft_living15, sqft_above, and bathrooms demonstrate the highest correlations with the price.

	Correlations	Features
3	0.875884	[sqft_living, sqft_above]
7	0.875884	[sqft_above, sqft_living]
2	0.764596	[sqft_living, grade]
5	0.764596	[grade, sqft_living]
6	0.758236	[grade, sqft_above]
8	0.758236	[sqft_above, grade]
4	0.758152	[sqft_living, sqft_living15]
9	0.758152	[sqft_living15, sqft_living]
0	0.753507	[bathrooms, sqft_living]
1	0.753507	[sqft_living, bathrooms]

Question 2: What features have the strongest correlations with other predicting variables?

Both sqft_living and sqft_above are highly correlated with price, however, only one of them can be included in a multiple regression model because they are also highly correlated with each other.
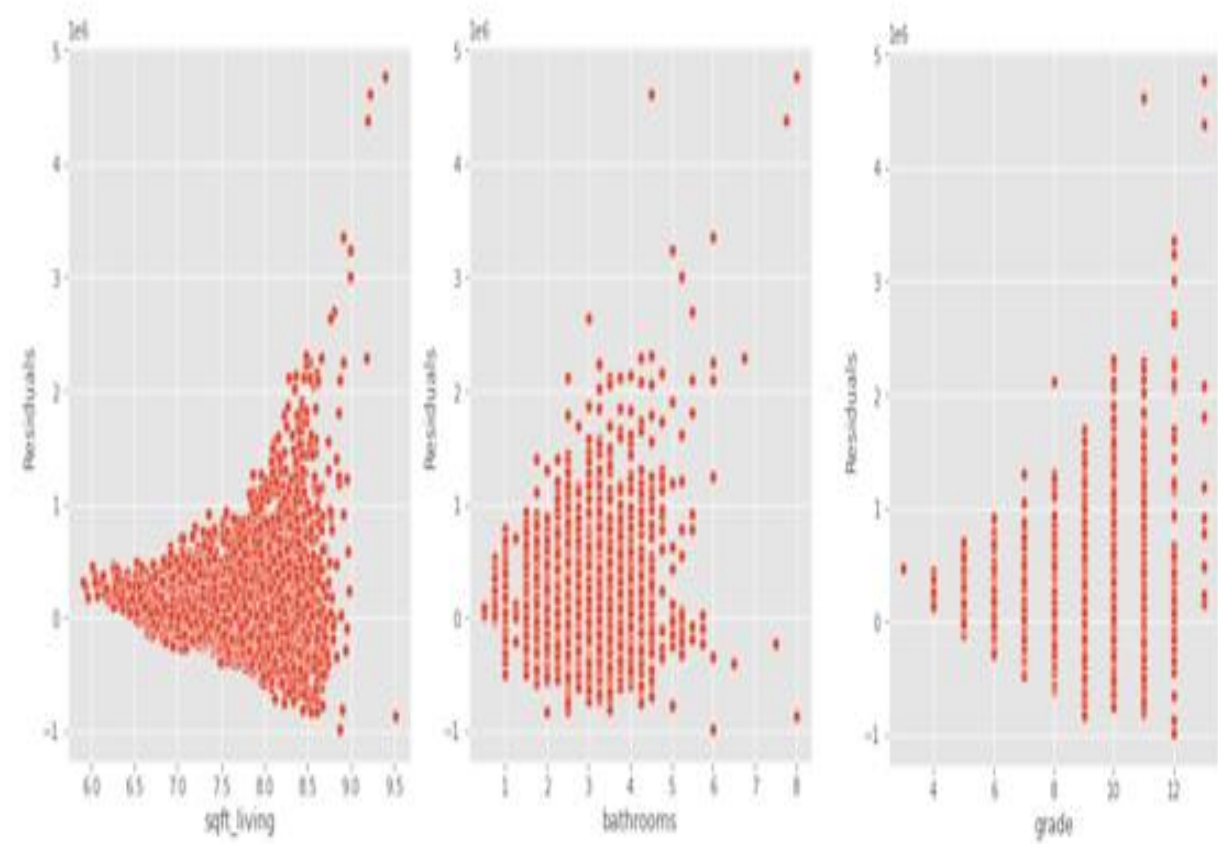
Question 3: What combinations of features is the best fit for price predictions?

- **Square footage.**

- **Number of bathrooms.**

- **The grade of the house.**

The above factors have a have statistically significant coefficients at a 95% confidence level. This means that these variables have a significant impact on the target variable, "price."

FINAL MODEL AND RESULTS

- **MODEL FIT**
a) R Squared ; 0.537
b) P Value < 0.05
- **MODEL COEFFIENTS**
a) Grade
b) Square foot living
c) Bathrooms

- This means the model can account for 54% of the variability in price.
- P Value < 0.05 means that the model is statistically significant.
- Our final model3 when compared to the baseline model, the R-squared increase from 50% to 54%

Recommendations

1.In King County, square footage, grade, and bathrooms are the most influential factors in determining a house's price.

2.Homeowners should prioritize expanding the square footage and enhancing the quality of construction to maximize their selling price.

3.Regularly retraining the model with the latest data, validating its predictions against actual sale prices, and refining it based on feedback are essential for maintaining the model's effectiveness.

Conclusion

1.Key factors in determining house prices in King County are square footage, grade, and bathrooms. Expanding square footage and improving construction quality can maximize selling prices.

2.Regular retraining of the model with updated data is necessary to capture market shifts and maintain effectiveness.

3.Thorough validation, testing, and analysis of the model's predictions against actual sale prices are crucial. Continuous refinement based on feedback and real-world performance insights should be implemented.

Thanks!

Do you have any questions?