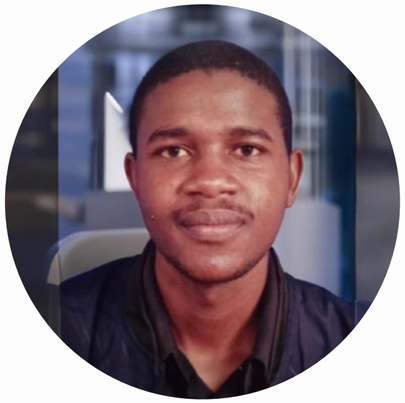
King County House Sales Prediction

GROUP - 18

GROUP 18 TEAM MEMBERS



Geoffrey Mwangi

Maureen Wanjeri

Veronicah Munyao

Victor Maina

Table Of Contents

PROJECT OVERVIEW

This project focuses on predicting house prices in King County, Washington, using regression modeling.

The goal is to provide real estate agents with accurate pricing recommendations for homeowners looking to sell their properties as well as homebuyers looking to get a fair deal. By setting the right price, agents can attract more buyers and ensure a quicker sale.

BUSINESS UNDERSTANDING

- By developing an accurate pricing model , homeowners will be able to estimate the value of their houses and sell them at competitive prices.

- Real estate agents will be able to advise their clients on pricing strategies and investors can identify potentially undervalued properties. This can lead to more efficient and profitable real estate transactions in King County.

OBJECTIVE

- To develop a regression model that predicts the sale price of homes based on various features such as square footage, number of bedrooms, bathrooms, location, and other relevant variables.

Data Understanding

- Missing values were identified.
- Key features impacting house prices were identified.
- Distributions of key features were visualized.
- Correlation analysis highlighted important relationships between features and house prices.
- Outliers in the data were identified.

Key Features used

1. Bedrooms
2. Bathrooms
3. Sqft_living
4. Sqft_lot

Correlation Analysis

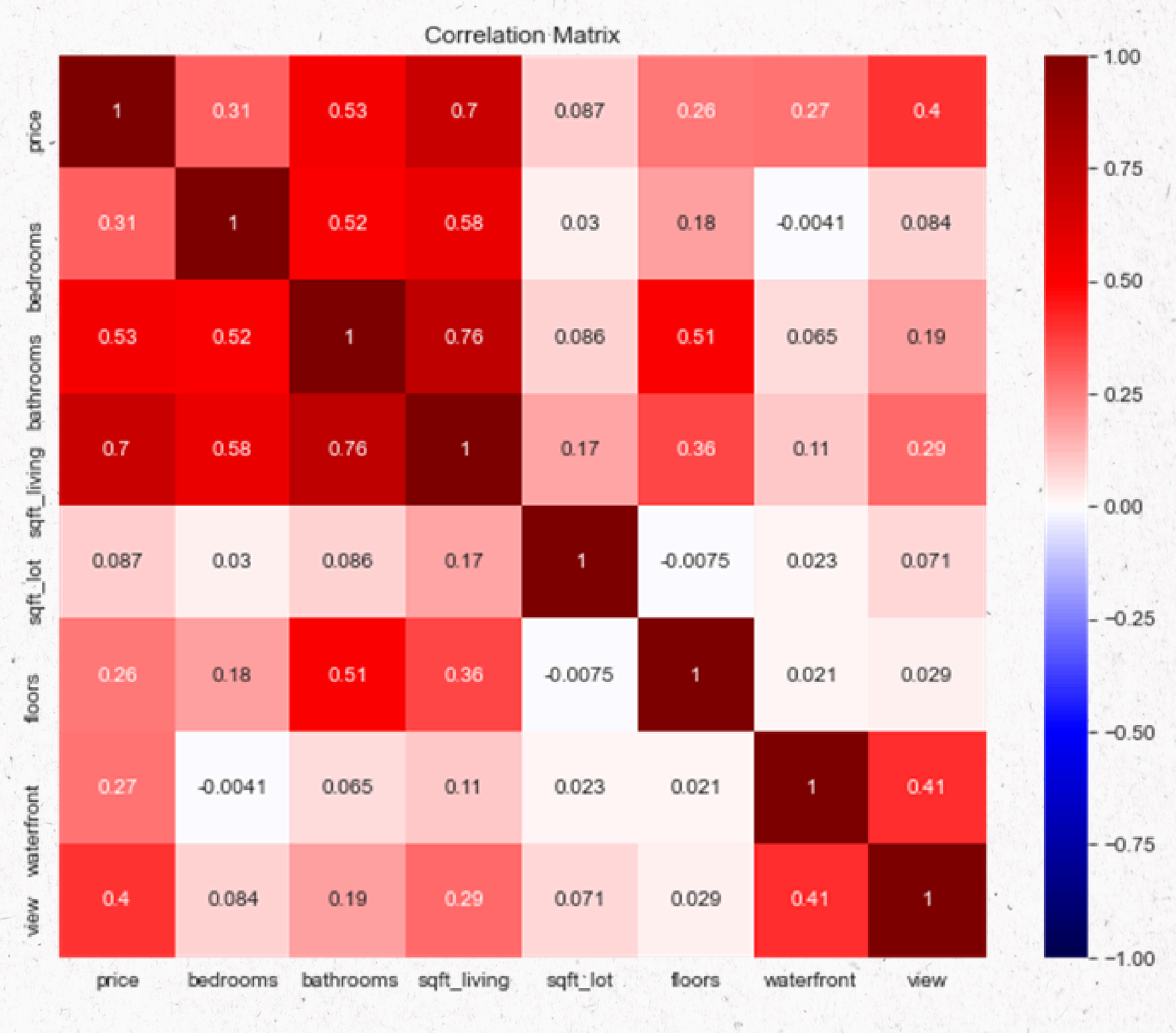
-Price is most strongly correlated with sqft_living and bathrooms.

-Bedrooms are strongly correlated with bathrooms and moderately correlated with sqft_living.

-Bathroom show a strong correlation with both sqft_living and bedrooms.

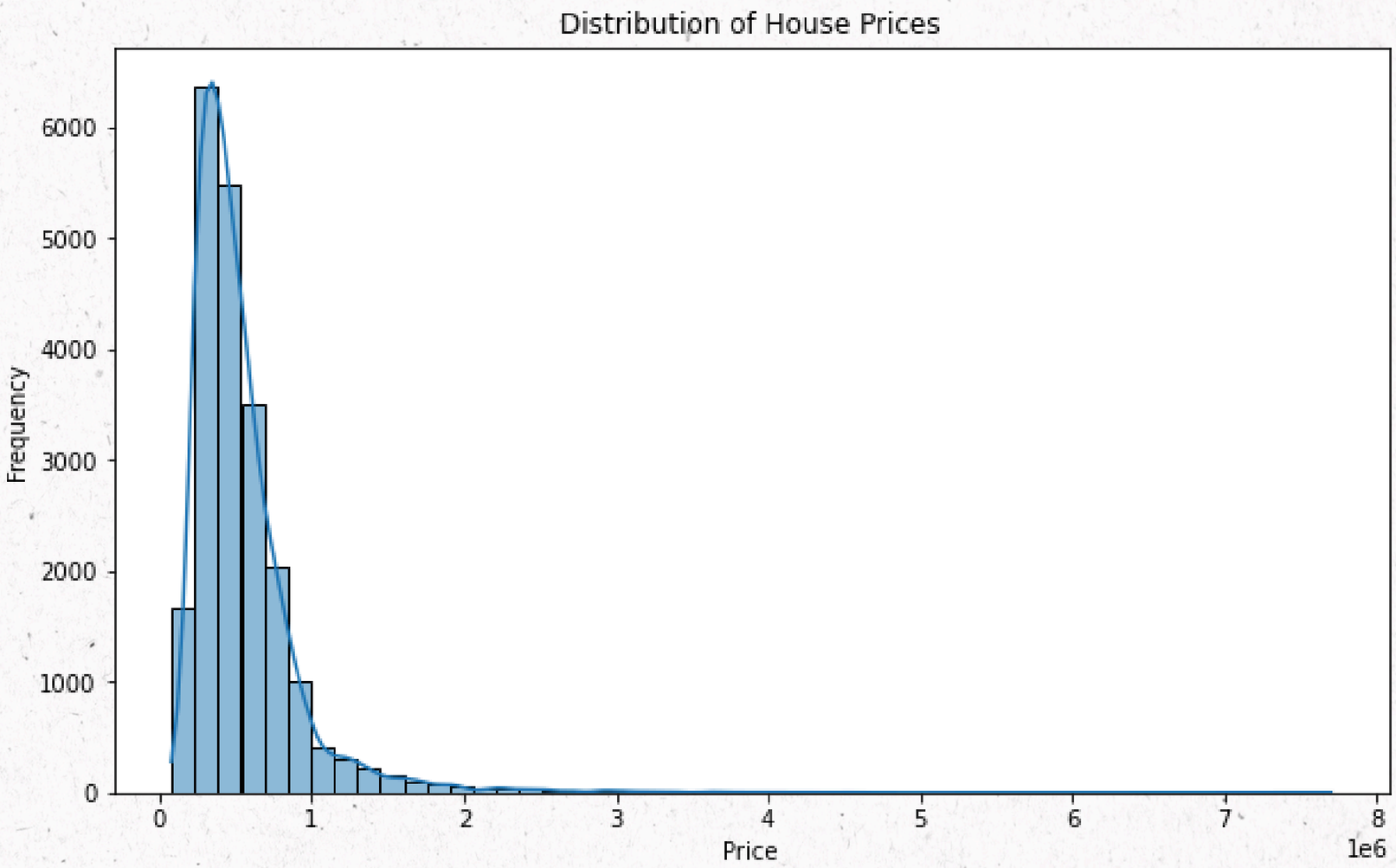
-Sqft_living has strong positive correlations with both bathrooms and price.

-These high correlations suggest that the size of the living area and the number of bathrooms are key factors influencing house prices and the number of bedrooms.



Distribution of House Prices

- Most houses are priced within a lower range, with a decreasing number of houses as the price increases.
- The distribution is heavily right-skewed, indicating that while high-priced houses exist, they are much less common.
- The presence of outliers on the higher end of the price range should be taken into account in further analyses, as they can significantly impact statistical measures such as mean and standard deviation.

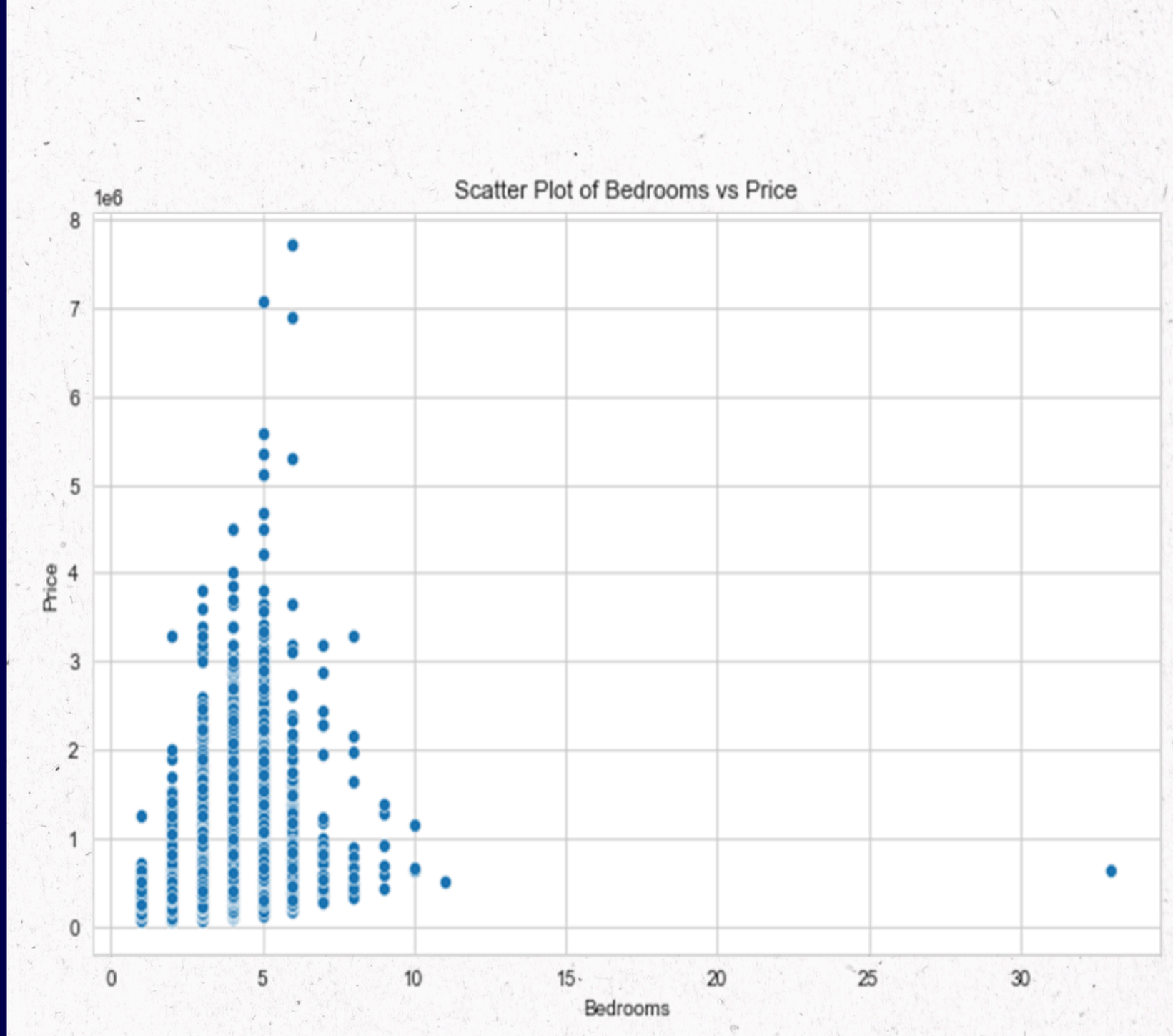


Visualizing Relationship between bedrooms and price

-Most houses have 3 to 6 bedrooms, with prices generally increasing up to 6 bedrooms beyond which prices do not show a clear trend.

-There are a few outliers with extremely high prices and an unusual property with over 30 bedrooms.

-Overall, the number of bedrooms is not strictly linearly related to house prices, indicating other significant influencing factors.



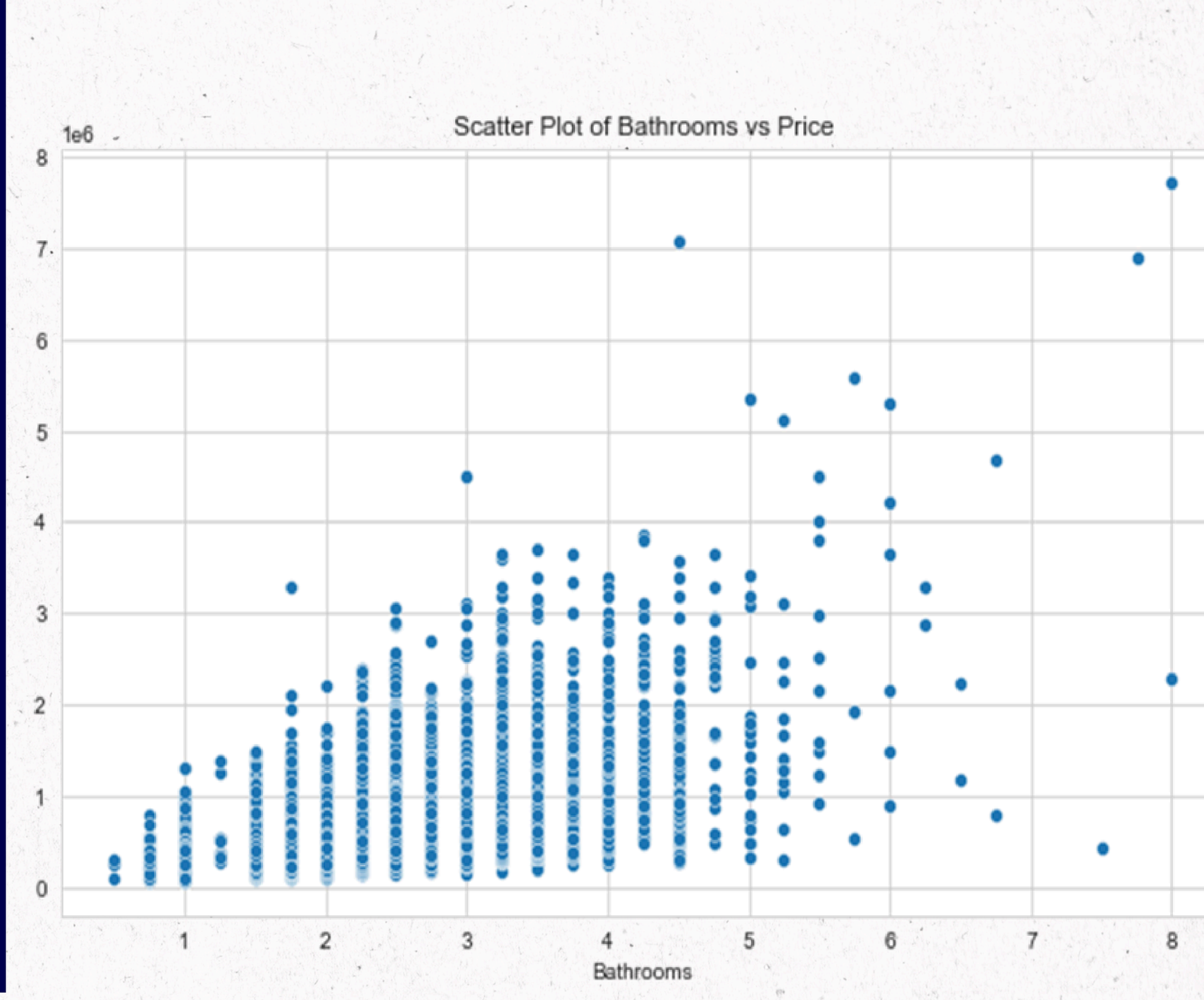
Visualizing relationship between bathrooms and price

-The scatter plot of bathrooms vs. prices shows a general positive correlation, indicating that homes with more bathrooms tend to be priced higher.

-Most homes have between 1 to 4 bathrooms, with prices generally up to 3 million.

-There is significant price variability within this range, suggesting other factors also influence prices.

-A few outliers with very high prices and many bathrooms indicate rare, high-value properties.



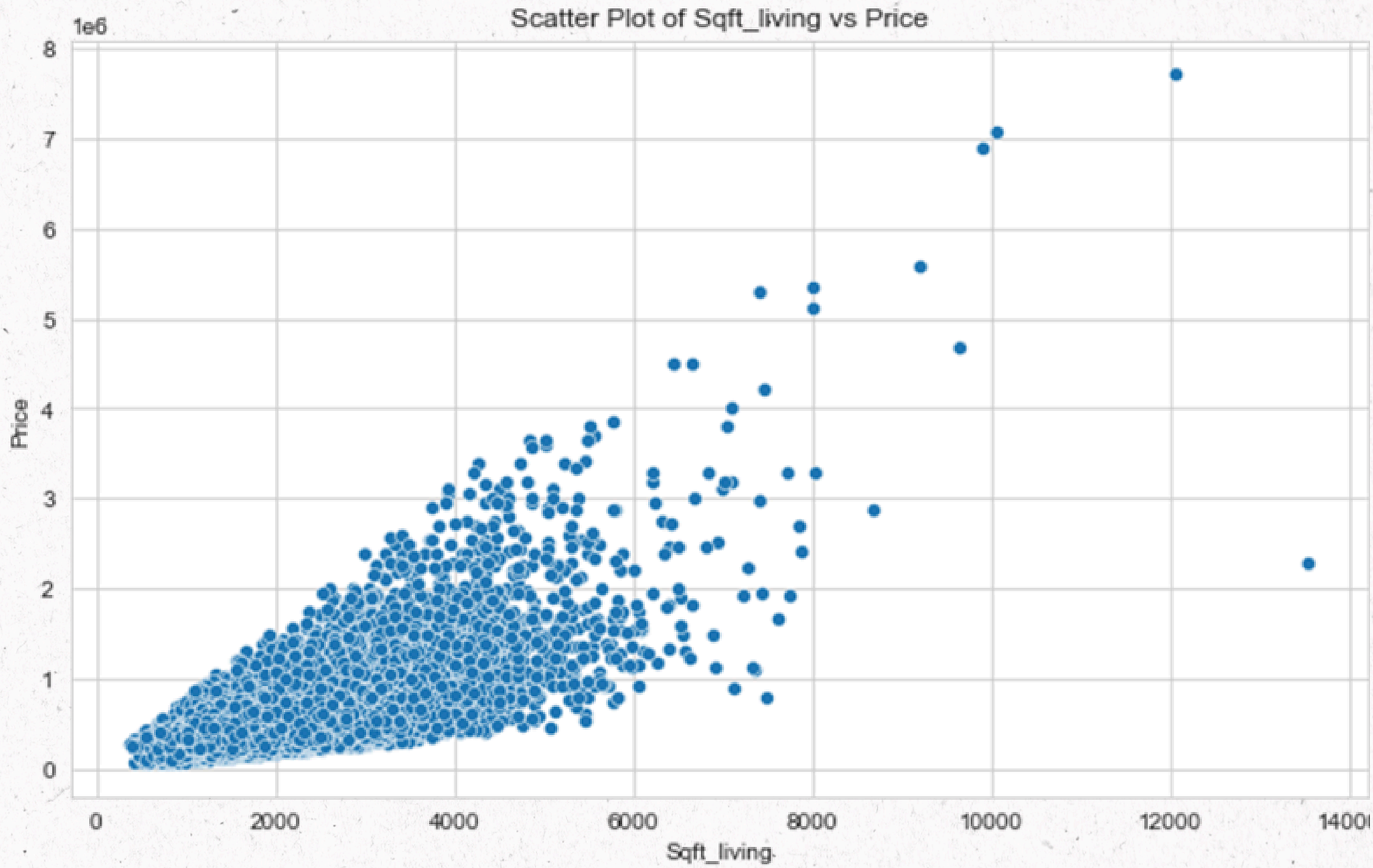
Visualizing relationship between sqft_living and price

-The scatter plot shows a strong positive correlation between the living area (in square feet) and the price, indicating that larger homes generally have higher prices.

-Most data points are concentrated between 1,000 and 4,000 square feet, with prices typically up to 2 million.

-There are some outliers with exceptionally large living areas (above 10,000 sqft) and high prices (above 5 million), highlighting rare, high-value properties.

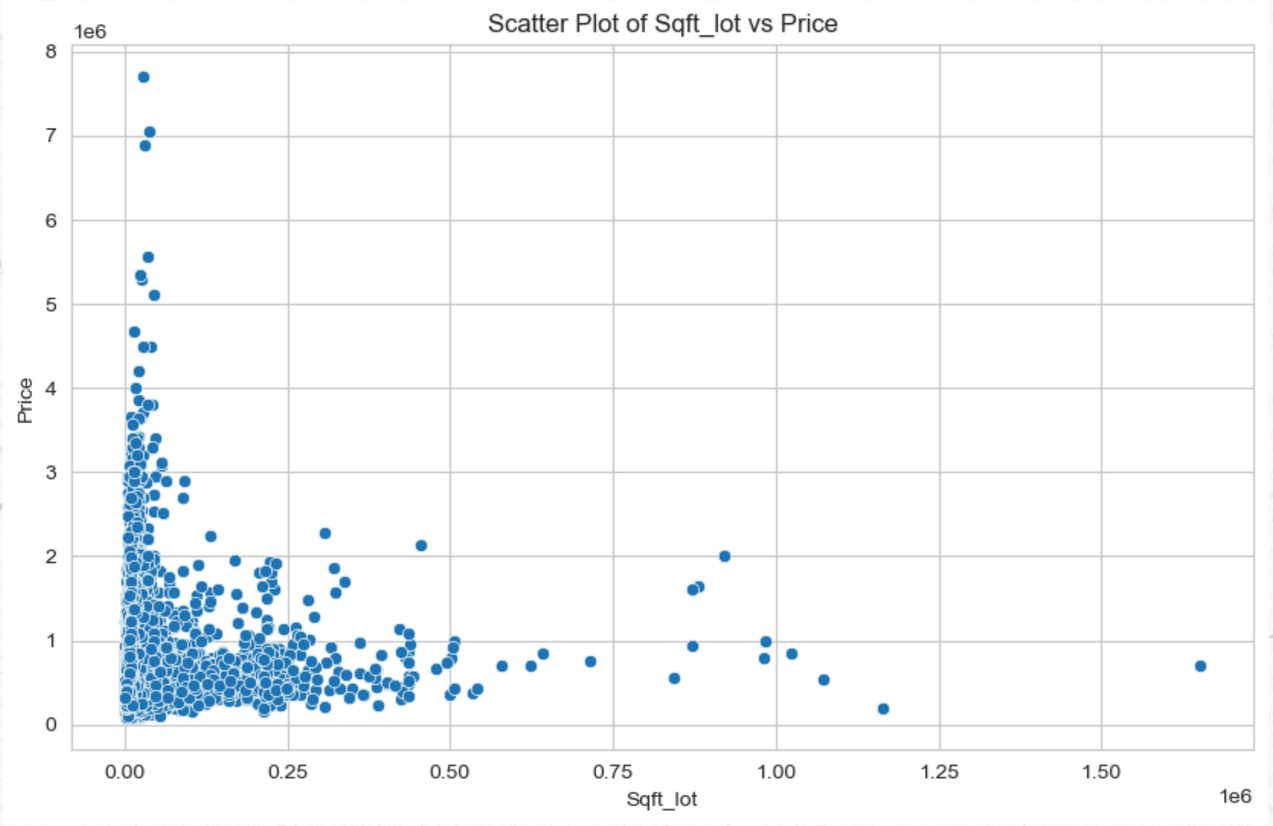
-This suggests that while living area is a key factor in determining price, other factors also play a significant role.



Visualizing relationship between sqft_lot and price

-The scatter plot shows a weak positive correlation between lot size and price, with most properties clustered below 0.25 million sqft_lot and price under 2 million.

-Beyond 0.25 million sqft, fewer data points and lower price variability suggest that lot size alone is not a strong determinant of property value.



Data Preparation

Data preparation significantly improved the quality and usability of our data, leading to more accurate predictions and better business decisions.

01 Handled missing values to prevent issues during analysis.

02 Transformed categorical data into a format that can be used by machine learning algorithms. (waterfront column)

03 Split the data into training and testing sets that ensures our models are robust and perform well on new, unseen data to :

- Prevent Overfitting
- Generalization to assess how well it generalizes to new, unseen data.

Modeling and Evaluation

01 Simple Linear

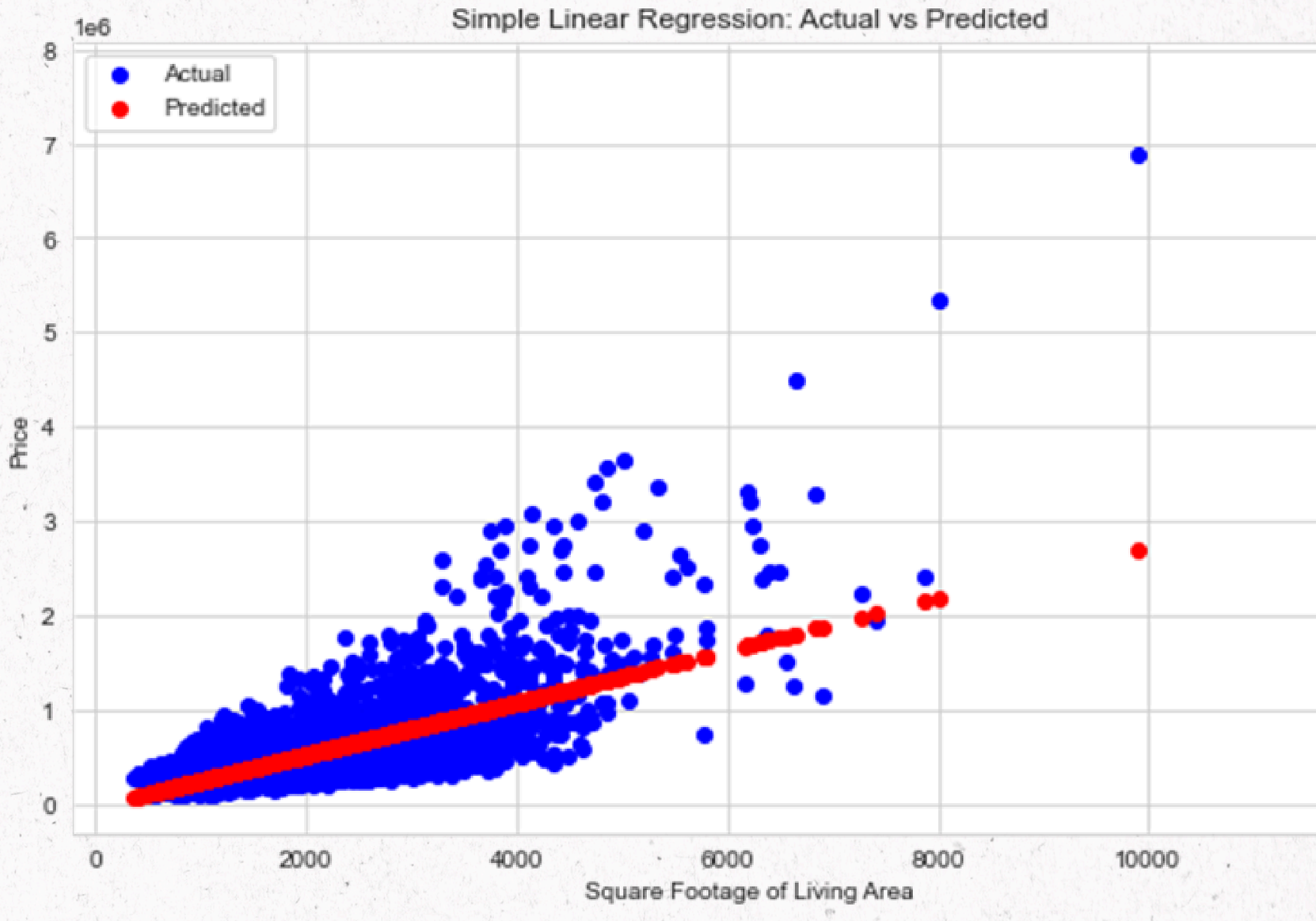
02 Multiple Linear

Simple Linear Regression

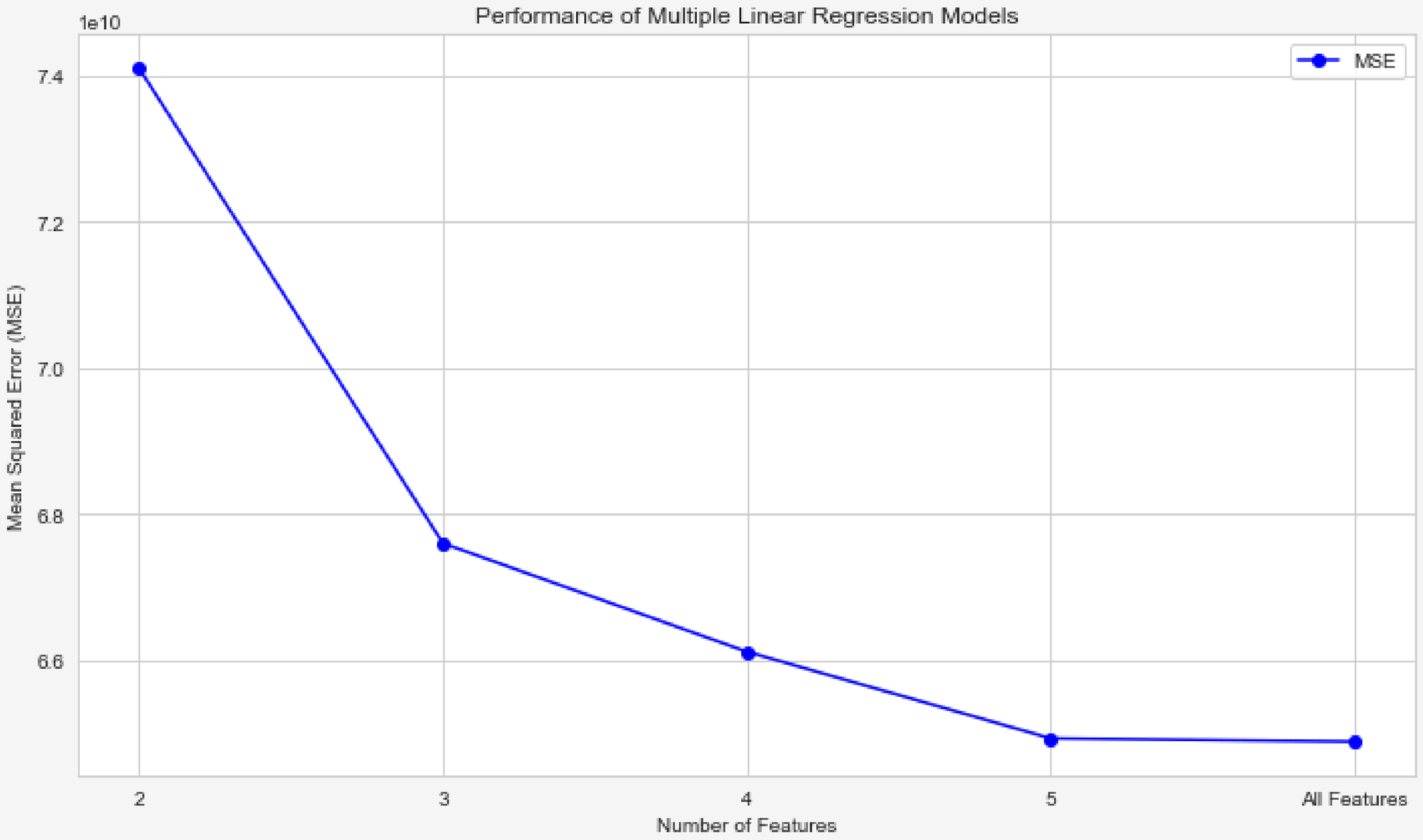
-The model is a good fit for lower values of square footage(up to around 4000_5000sqft), predicted prices(red dots) are relatively close to the actual pricees(blue dots), indicating that the model is performing reasonably well in this range.

-Model is a poor fit for higher square footage as there is a noticeable deviation between the actual prices and the predicted prices. Model tends to underpredict prices for larger houses as evidenced by the blue dots being significantly higher than the red dots.

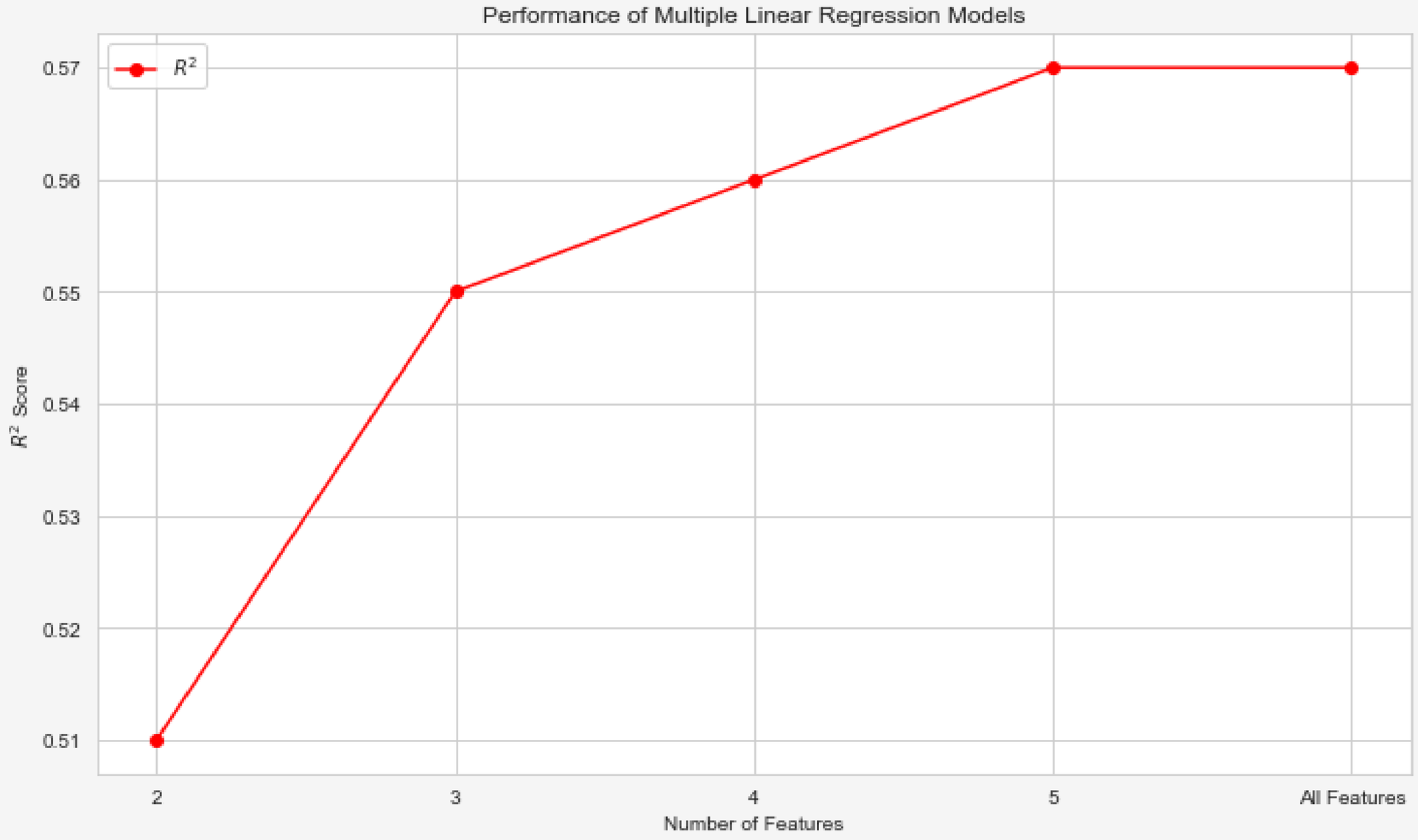
-There are some significant outliers where the actual prices are much higher than the predicted prices, especially for houses with very high square footage(above 8000 sqft) suggesting that the linear model might not be capturing the complexity of the relationship for these larger properties.



A lower MSE indicates that the Multiple Linear Regression model has a better fit to the data and makes more accurate predictions.



A higher R^2 score (0.576 compared to 0.503) suggests that the Multiple Linear Regression model explains a greater proportion of the variance in the dependent variable, making it a more reliable model for predicting outcomes.



MODEL EVALUATION

Model 1

Using only sqft_living and bathrooms, we explain 51% of the variability in prices with a relatively high MSE.

Model 2

Adding view improves the model to explain 55% of the variability and reduces the MSE.

Model 3

Further adding bedrooms improves the model slightly more, explaining 56% of the variability and reducing the MSE.

Model 4

Including waterfront improves the model slightly, explaining 57% of the variability and reducing the MSE further.

Model 5

Including all features (adding floors) does not significantly improve the model over

Modelling Conclusion

- **Incremental Improvement:** Each step of adding a new feature generally improves the model, reducing the MSE and increasing R^2.

- **Best Model**: Model 4 and Model 5 both explain 57% of the variability in house prices with a slight difference in MSE. Therefore, Model 4 can be considered optimal for its simplicity and performance.

Business Conclusion

01 Recommendations

02 Next Steps

RECOMMENDATIONS

1 Focus on Living space

-Emphasize the total living space in marketing materials. Houses with larger living areas are highly valued.

2 Enhance property grade and condition

-Invest in improvements that can increase the overall grade and condition of the property

3 Capitalize on waterfront and views

-If the property has a waterfront view or high_quality view, we make sure to highlight this feature prominently

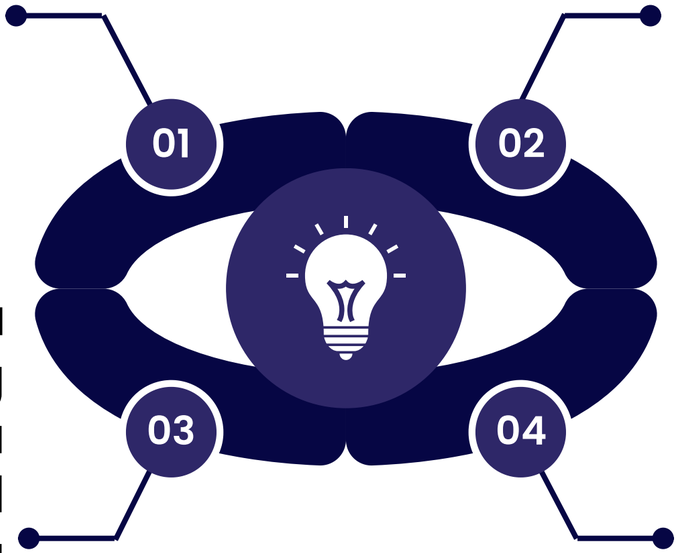
4 Promote bathroom additions

-Properties with more bathrooms tend to have higher prices

Next Steps

- **Incorporating more advanced machine learning techniques**: While linear regression and decision trees are commonly used for house price prediction, there are many more advanced techniques that could be explored, such as neural networks, gradient boosting, and support vector machines.

- **Incorporating external data sources:** The House Sales in King County, USA dataset contains a limited set of features, and incorporating external data sources such as weather data, crime rates, and school quality ratings could provide additional insights into the factors that influence house prices.



- **Exploring the impact of different types of features:** While the House Sales in King County, USA dataset contains a wide range of features, there may be other types of features that could be more important for predicting house prices, such as features related to the age and income of the homeowners.

- **Investigating the impact of different geographic regions:** While the House Sales in King County, USA dataset focuses on houses in one specific geographic region, it would be interesting to investigate whether similar predictive models could be developed for other regions, or whether there are unique factors that influence house prices in different regions

THANK YOU