KINGS COUNTY HOUSING PROJECT

Randell Mwania

PROBLEM

LINEAR REGRESSION ANALYSIS OF HOUSE SALES IN A NORTHWESTERN COUNTY

The objective of this project is to build a model that can predict house prices based on the features of the house using the available King's County house-selling records.
The model can be used by both sellers and buyers in their business decisions. Sellers can predict the selling price of their house and determine if any renovations are necessary before selling. Buyers can receive suggestions on the type of house they can afford based on their budget.

The following objectives are set to achieve the final goal:

1. Analyze and clean the data by handling meaningless or null values.
2. Remove features that do not contribute to the house price.
3. Identify highly correlated features and potentially remove redundant ones.
4. Build a linear regression model.
5. Evaluate the impact of different features on house prices.

PROJECT OVERVIEW

- Dataset: We have worked with a comprehensive dataset that contains valuable information on house prices and various features.
- Objective: Our main objective was to develop a robust predictive model that accurately estimates house prices based on the provided features.
- Importance: Accurate house price predictions are essential for real estate professionals, investors, and homebuyers to make informed decisions and gain insights into market trends.

BUSINESS UNDERSTANDING

- Understanding the stakeholders: We aimed to cater to the needs of real estate professionals, investors, and homebuyers who rely on accurate house price predictions for their decision-making processes.

Key questions:

- What are the primary factors influencing house prices?
- How effectively can we estimate house prices based on the provided features?
- What recommendations can we offer to stakeholders based on our analysis?

BUSINESS UNDERSTANDING

Understanding the stakeholders:

- I aimed to cater to the needs of real estate professionals, investors, and homebuyers who rely on accurate house price predictions for their decision-making processes.

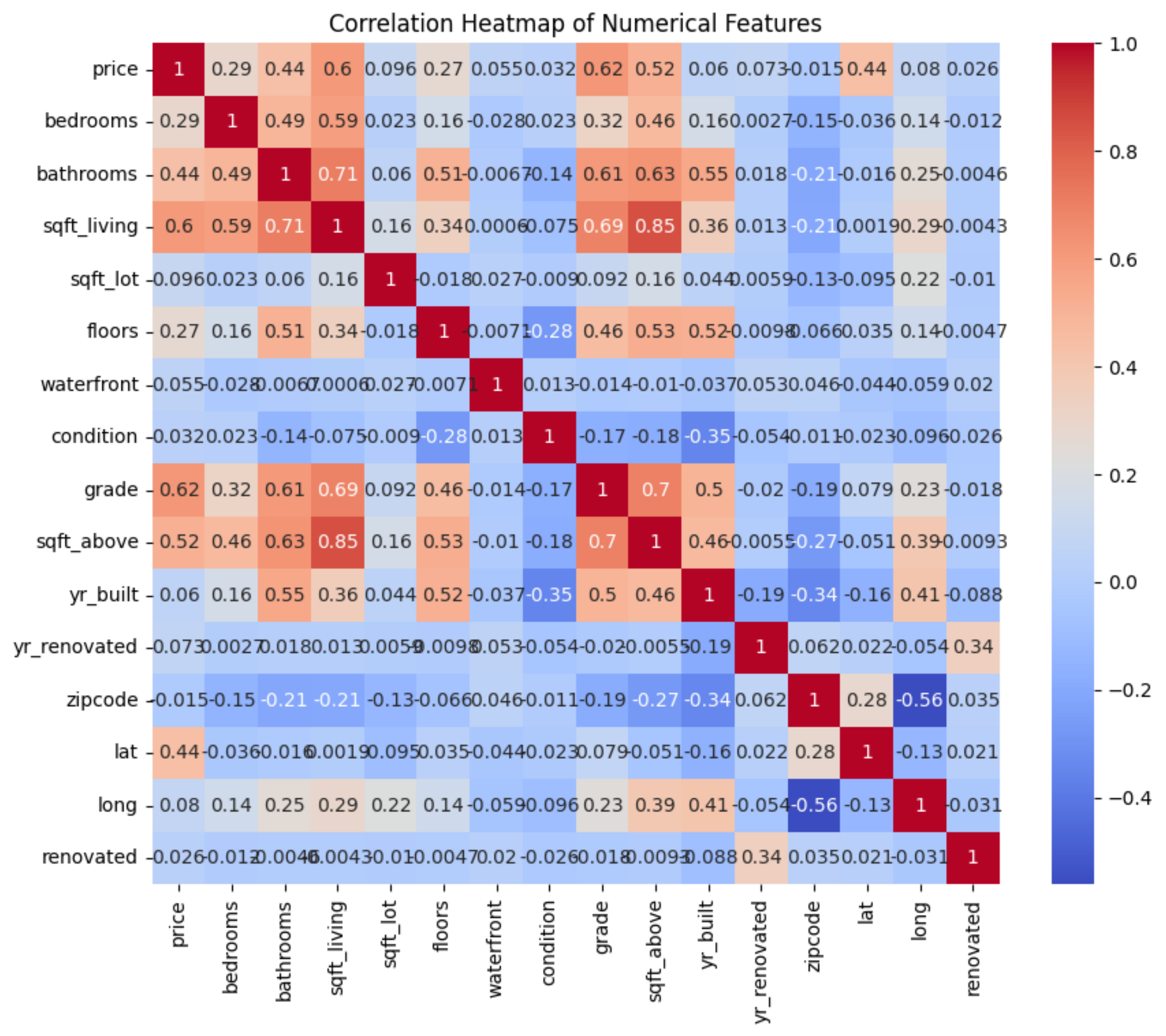
Key questions:

- What are the primary factors influencing house prices?
- How effectively can we estimate house prices based on the provided features?
- What recommendations can we offer to stakeholders based on our analysis?

DATA UNDERSTANDING

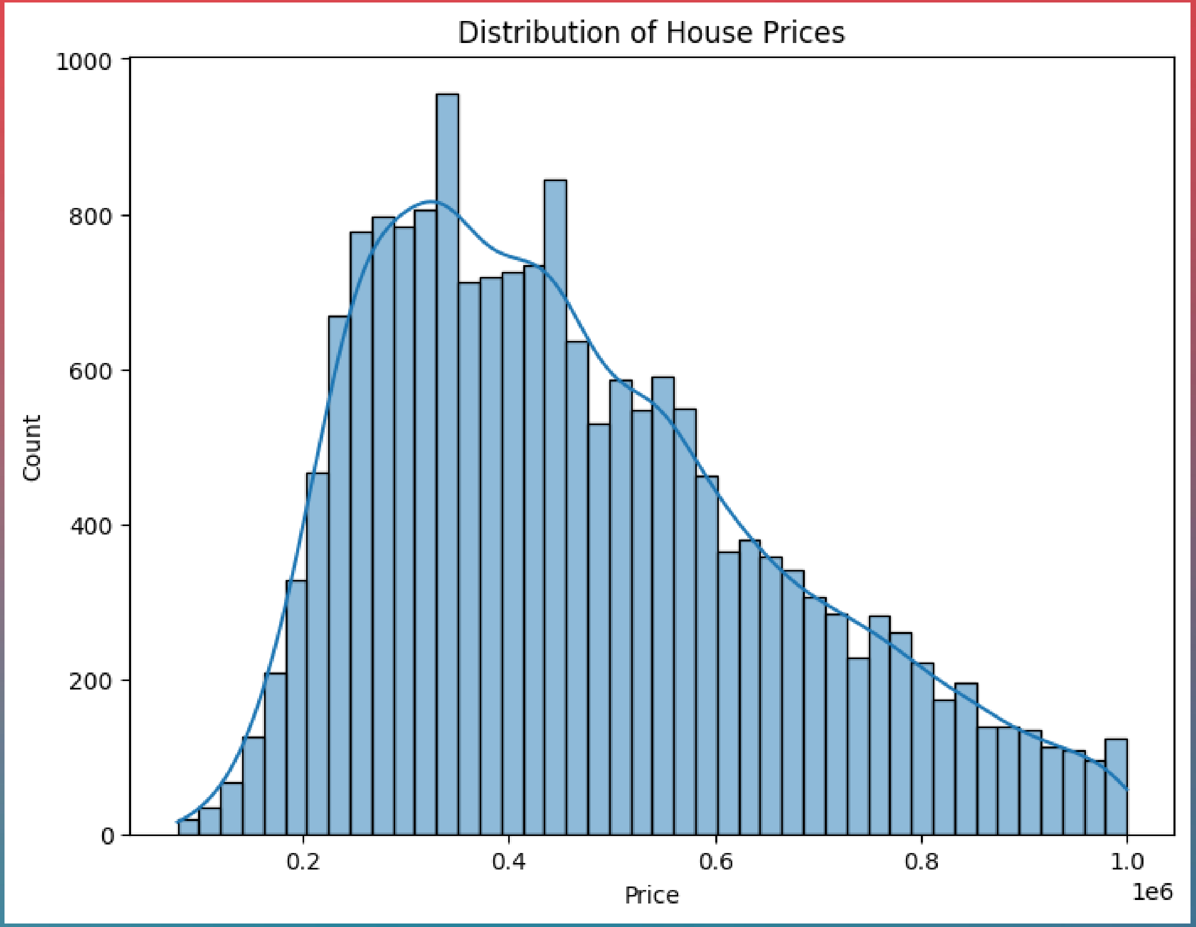
- Dataset description: Our dataset comprises comprehensive information about house characteristics, including bedrooms, bathrooms, square footage, floors, year built, and more.
- Thorough data preparation: To ensure data quality, I conducted various data preparation steps. These included removing unnecessary columns, addressing missing values, converting data types, and creating new features.

CORRELATION HEATMAP OF NUMERICAL FEATURES TEN MOVIES BY NUMBER OF VOTES



This visualization helps to understand the relationships between the numerical features in the dataset. Positive correlations are represented by warmer colours (red) and negative correlations by cooler colours (blue). The intensity of the colour indicates the strength of the correlation.

DISTRIBUTION OF HOUSE PRICES



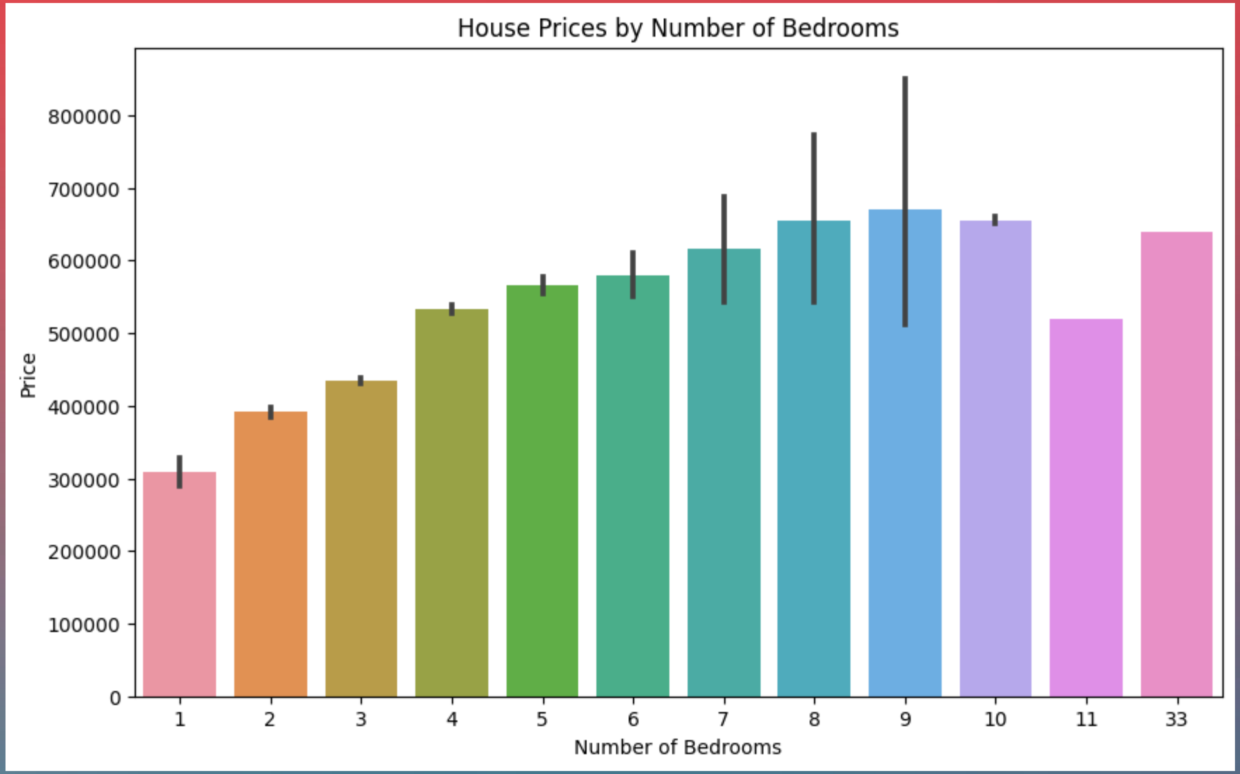
The histogram shown gives us a clear picture of how house prices are distributed in the dataset. On the x-axis, we have the different price ranges for the houses, while the y-axis represents the number of houses falling within each range. To provide a smoother representation of the data, a kernel density estimate (KDE) curve is overlaid on the histogram.

Main observations we can make from the histogram:

- The majority of house prices fall between $200,000 and $400,000, with the highest concentration around $300,000. This means that most houses in the dataset have prices within this range.
- As we move away from this price range, the number of houses gradually decreases. This suggests that houses with prices outside the $200,000 to $400,000 range are relatively less common in the dataset.
- The distribution of house prices appears to be slightly skewed to the right. In other words, there are more houses with lower prices compared to higher prices. This observation is supported by the KDE curve, which shows a gradual decline in density as prices increase.

Overall, this visualization provides us with valuable insights into the distribution of house prices in the dataset. It helps us understand the range of prices and where the majority of houses fall within that range.

HOUSE PRICES BY NUMBER OF BEDROOMS



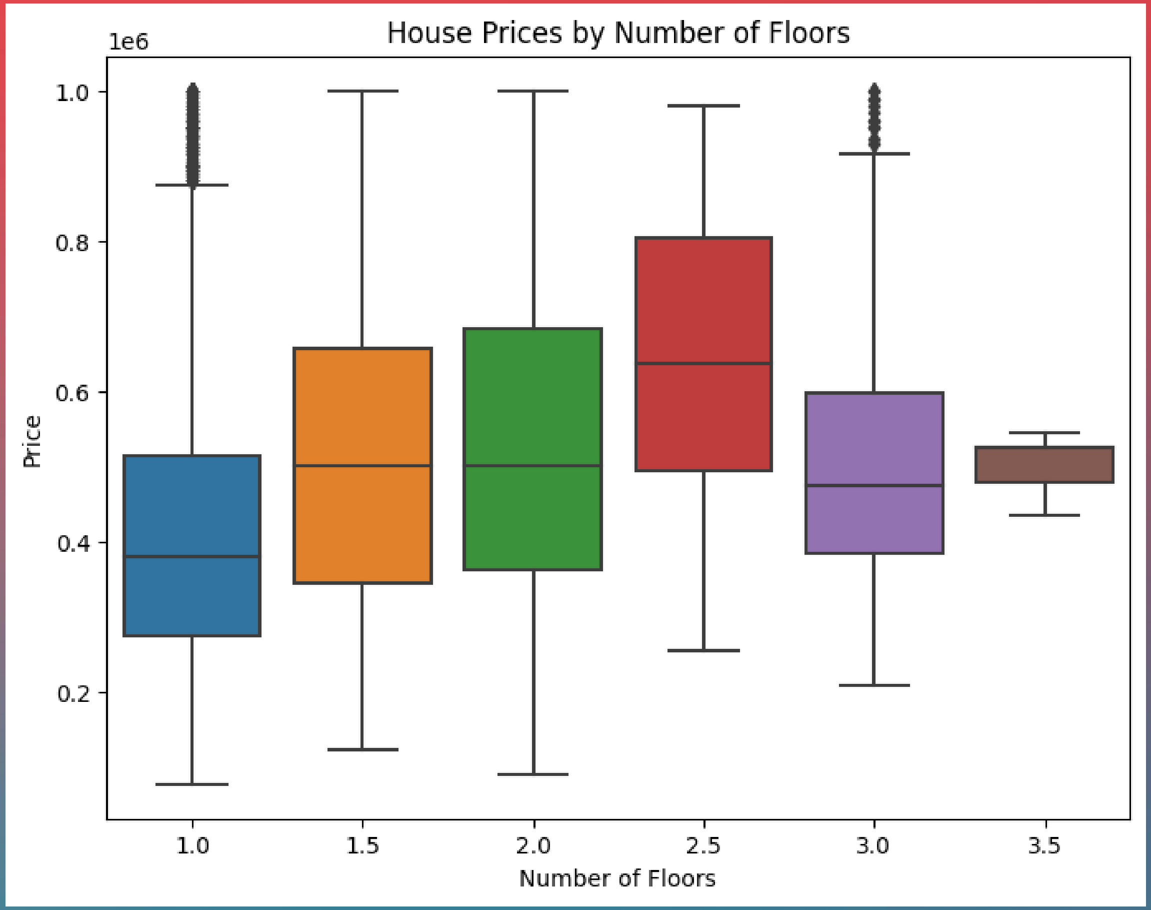
The bar plot shows the relationship between the number of bedrooms in a house and its price. The x-axis represents the number of bedrooms, while the y-axis represents the price.

From the plot, we can observe that houses with a higher number of bedrooms tend to have higher prices. This indicates that the number of bedrooms is a significant factor in determining the price of a house.

The plot also allows us to see the variation in prices for houses with different numbers of bedrooms. For example, houses with 4 bedrooms have a wider range of prices compared to houses with 1 or 2 bedrooms.

Overall, this plot provides a visual representation of the relationship between the number of bedrooms and house prices, allowing us to easily compare and analyze the data.

HOUSE PRICES BY NUMBER OF FLOORS



The boxplot provided above offers insights into the association between the number of floors in a house and its corresponding price. The x-axis represents the number of floors, while the y-axis represents the price.

Here are the observations we can make from the boxplot:

1. The majority of houses in the dataset have either one or two floors, as indicated by the taller boxes located towards the bottom of the plot.

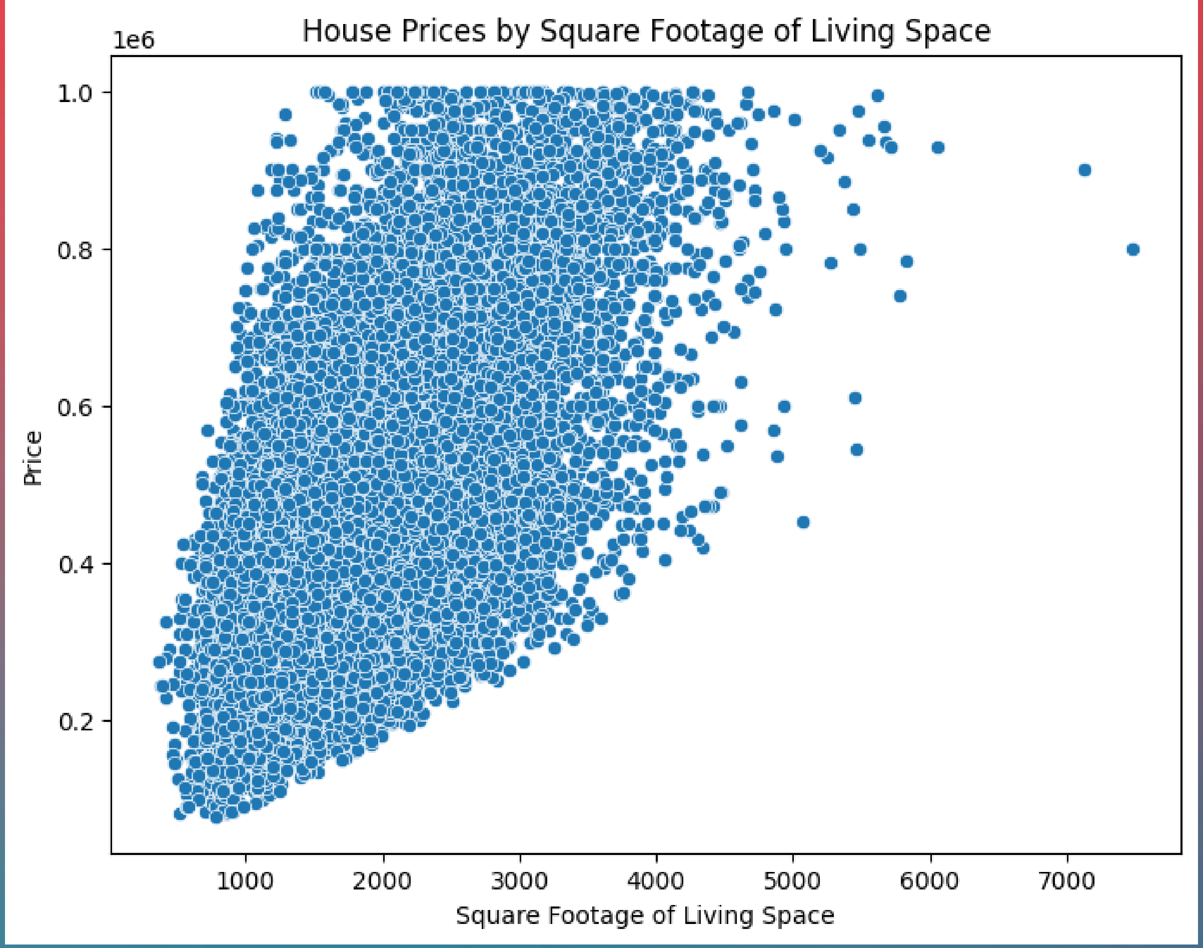
2. Houses with three or more floors are relatively uncommon, as demonstrated by the shorter boxes positioned towards the top of the plot.

3. In general, houses with a greater number of floors tend to have higher prices. This can be inferred from the increasing median price as the number of floors increases.

4. There are a few outliers present in the dataset, which are represented by individual points located outside the whiskers of the boxes. These outliers indicate houses with substantially higher prices compared to others with a similar number of floors.

To conclude, the boxplot suggests a positive correlation between the number of floors and the price of a house. However, it is important to acknowledge that other factors may also influence the price, and further analysis would be necessary to determine the exact relationship.

HOUSE PRICES BY SQUARE FOOTAGE OF LIVING SPACE



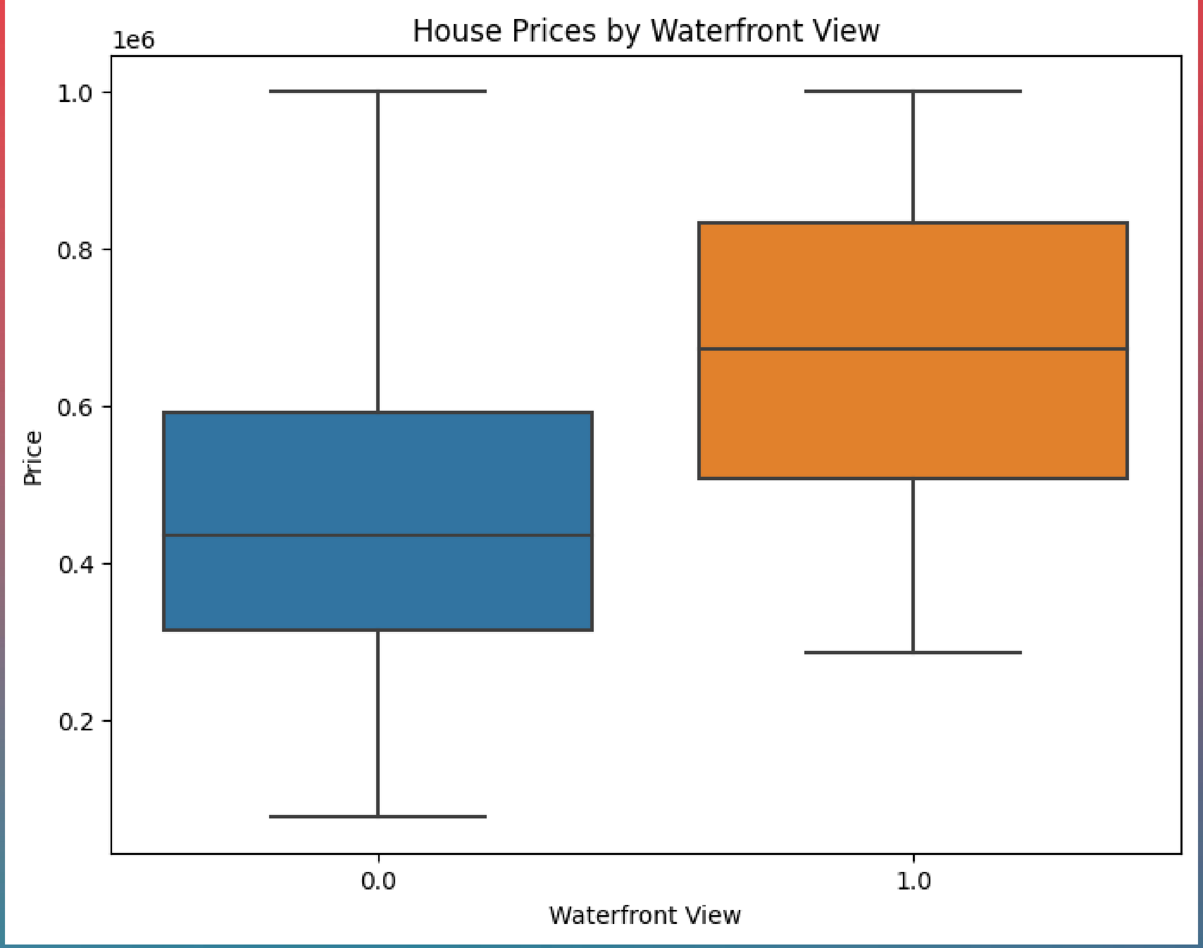
The scatter plot above shows the relationship between the square footage of living space and the price of houses. As the square footage of living space increases, there is a general trend of higher prices.

However, there is also some variation in prices for houses with similar square footage.

This suggests that other factors, such as location, condition, and amenities, may also influence house prices.

Overall, this plot provides a visual representation of the positive correlation between the square footage of living space and house prices.

HOUSE PRICES BY WATERFRONT VIEW

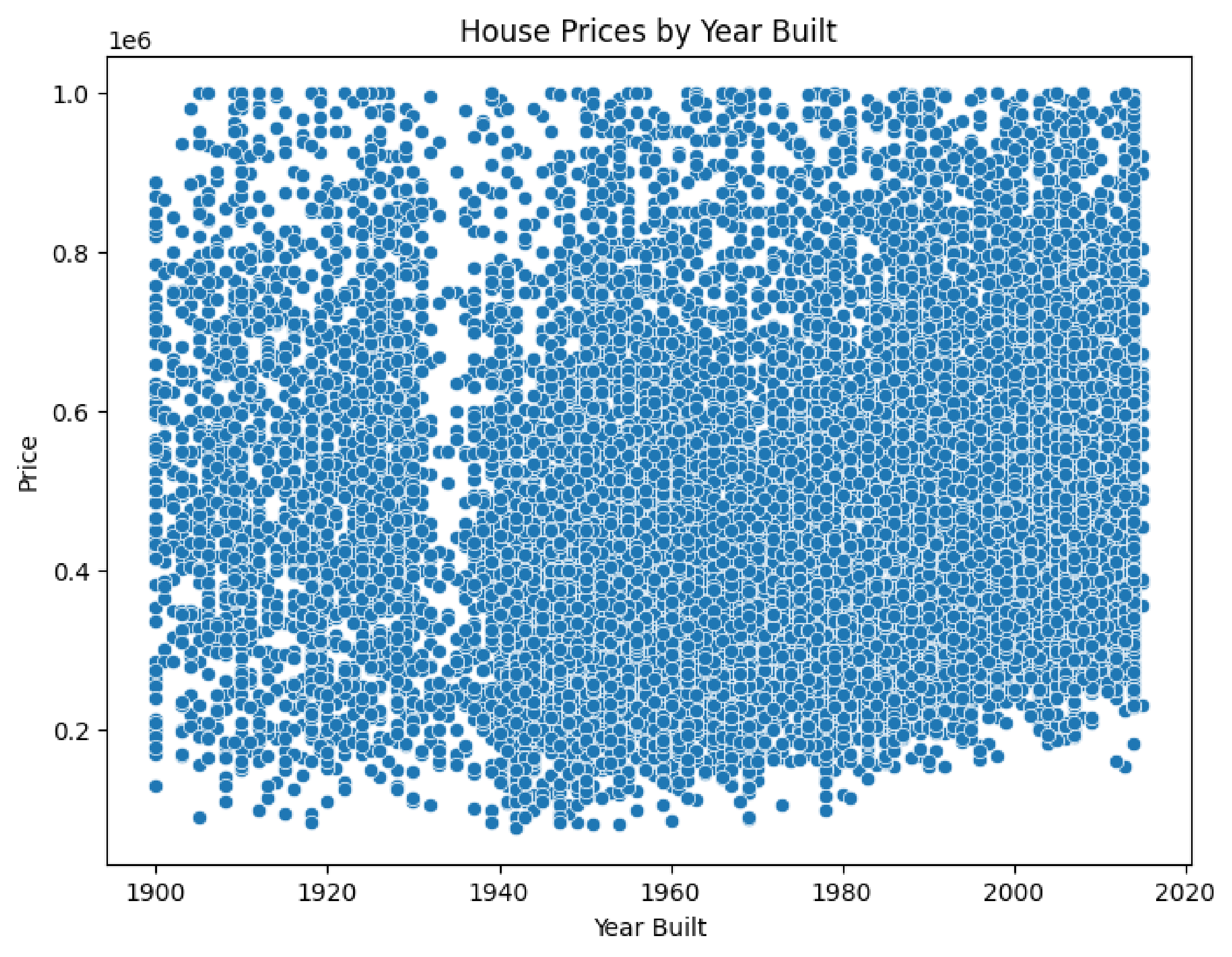


The boxplot shown above provides insights into how house prices are distributed based on whether or not they have a waterfront view. By examining the plot, it becomes clear that houses with a waterfront view tend to be priced higher compared to those without. This conclusion is supported by the observation that the median price for houses with a waterfront view is greater than the median price for houses without.

Furthermore, the range of prices for houses with a waterfront view is wider, indicating greater variability in prices. This suggests that factors beyond just the presence of a waterfront view might also be influencing the prices of these houses.

In summary, this boxplot serves as a visual representation of the connection between having a waterfront view and house prices, highlighting that houses with a waterfront view generally command higher prices.

HOUSE PRICES BY YEAR BUILT



The scatterplot analysis reveals that there is no clear linear relationship between the year a house was built and its corresponding price.

However, several things can be observed.

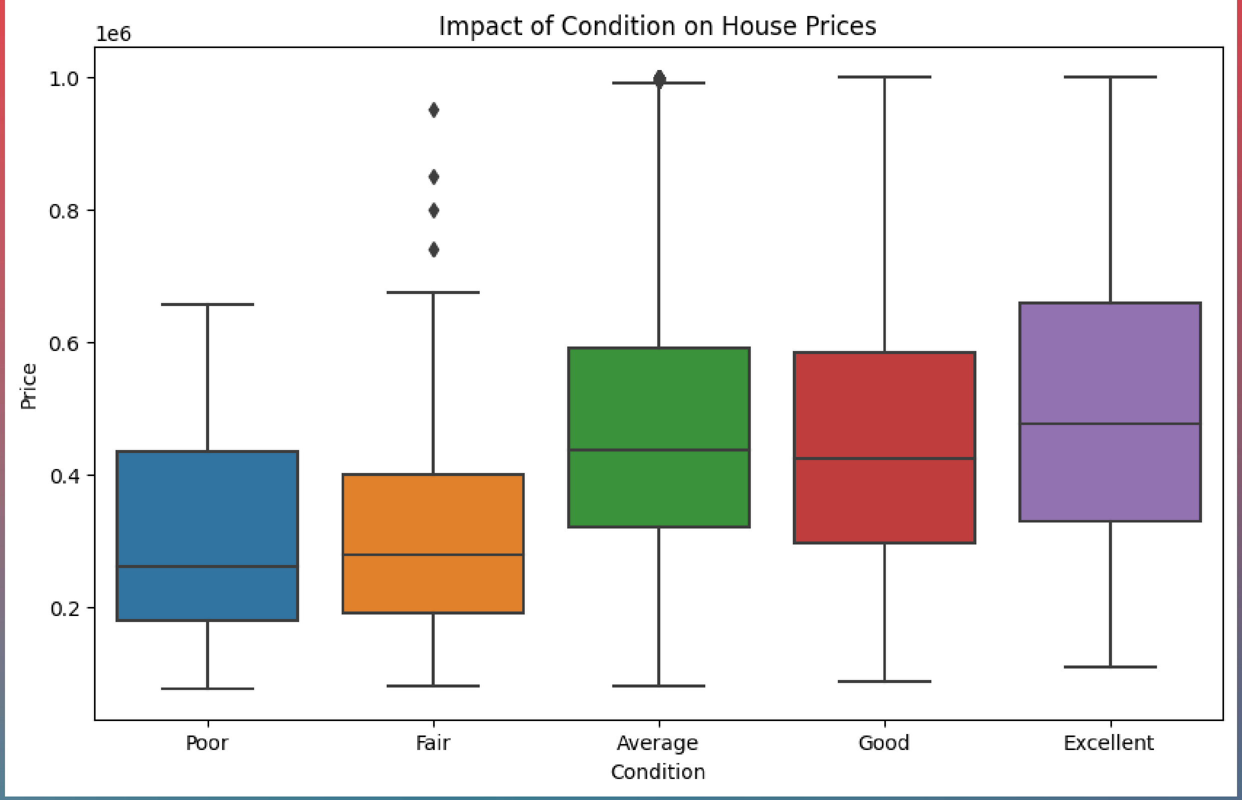
Houses built during the mid-1900s (around 1950-1970) exhibit a wide range of prices, indicating significant variation in value within this time period.

Some older houses built in the early 1900s (around 1900-1920) have relatively high prices, likely due to their historical or architectural significance.

Additionally, houses built in recent years (around 2010-2015) also command high prices, possibly attributed to modern design, updated amenities, or desirable locations.

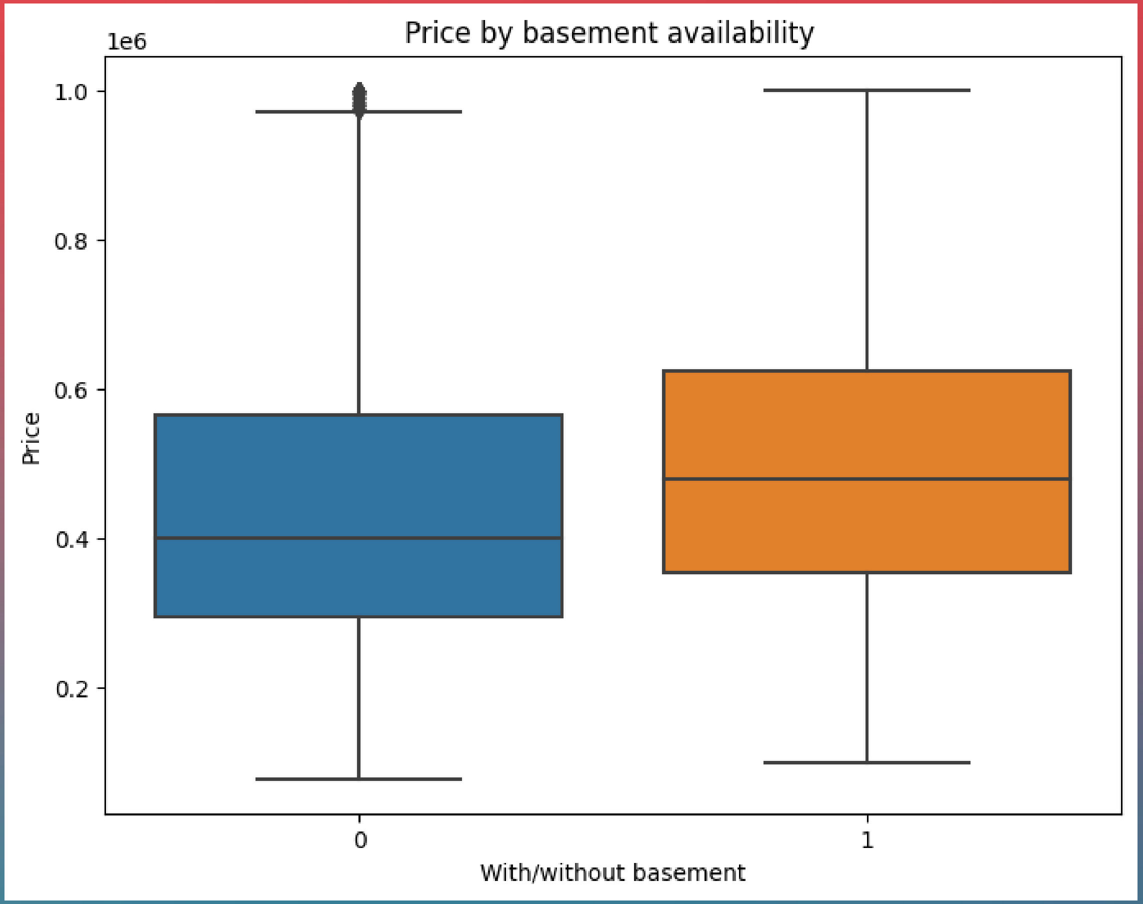
While the scatterplot does not demonstrate a strong correlation between year built and price, it provides insights into the relationship between these variables, highlighting the influence of historical value, architectural significance, and contemporary features on house prices.

IMPACT OF CONDITION ON HOUSE PRICES



- Condition and Price: Houses in poor condition generally have the lowest median price, indicating that their condition significantly impacts their value. On the other hand, houses in excellent condition have the highest median price, indicating that well-maintained properties command higher prices.
- Price Variation: The interquartile range (IQR) for houses in poor condition is relatively small, indicating less variation in prices. This suggests that houses in poor condition tend to have a more consistent pricing pattern. In contrast, the IQR for houses in excellent condition is relatively large, indicating a wider range of prices. This indicates more variability in prices among houses in excellent condition.
- Outliers: Both condition categories have outliers, represented by dots outside the whiskers of the boxes. These outliers signify houses with unusually high or low prices within their respective condition categories. These outliers may be influenced by factors beyond condition alone, such as unique features, location, or other market dynamics.

PRICE BY BASEMENT AVAILABILITY

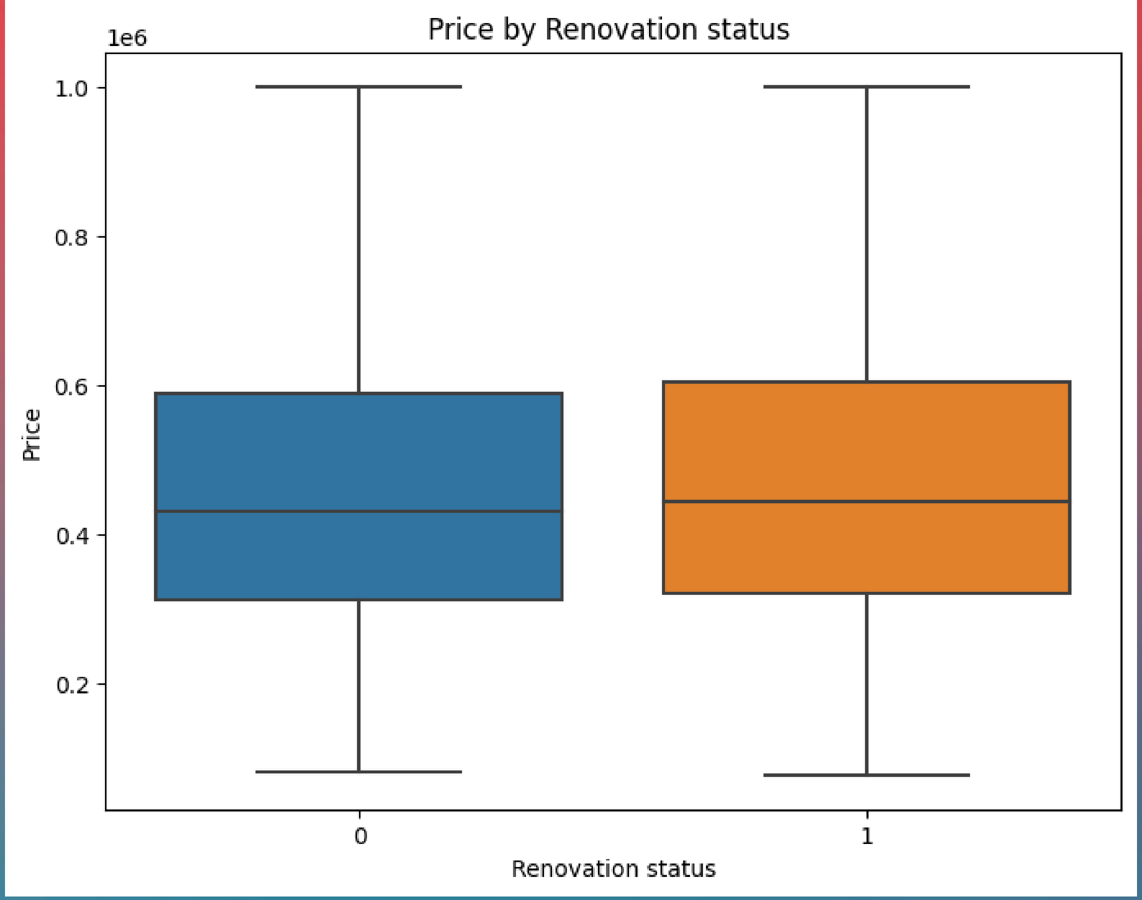


From the boxplot, we can observe that properties with a basement tend to have higher prices compared to those without a basement.

The boxplot also shows that properties without a basement have a few outliers with higher prices, indicating that there are some exceptional cases where properties without a basement can still be expensive.

In conclusion, the presence of a basement generally contributes to higher property prices, although there are exceptions where properties without a basement can still have high prices.

PRICE BY RENOVATION STATUS



Median Price: The median price of properties that have been renovated (renovated=1) is higher than the median price of properties without any renovation (renovated=0). This indicates that renovated properties generally command higher prices compared to properties that have not undergone any renovation.

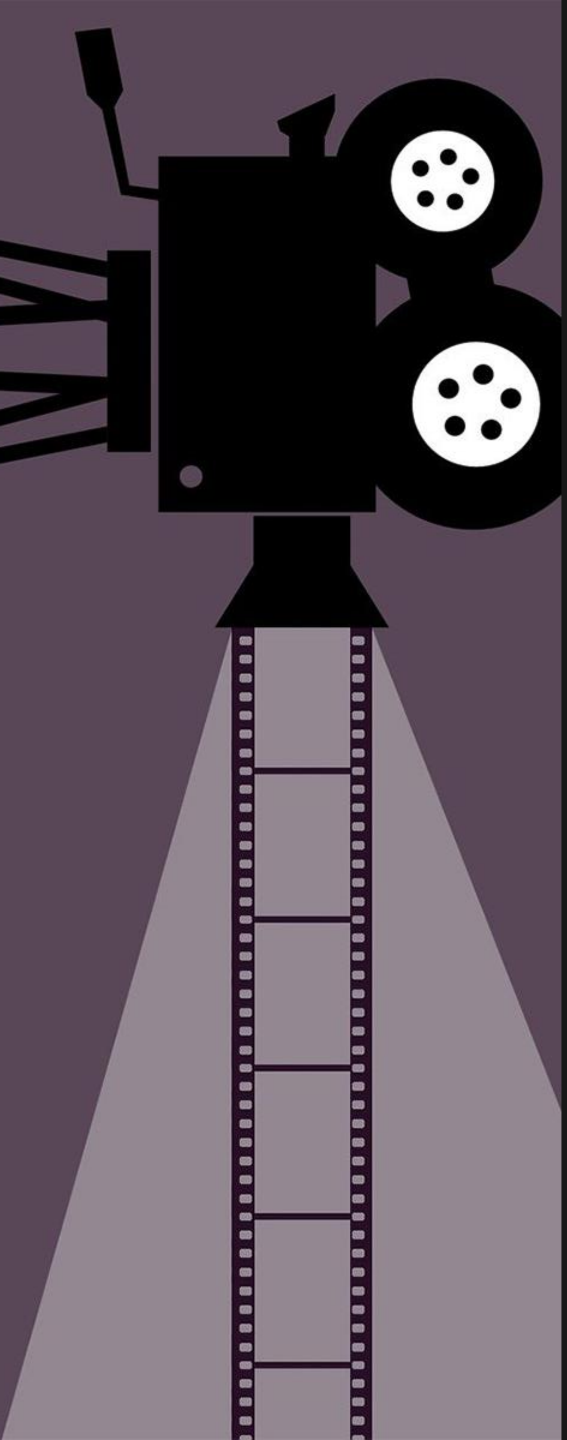
Price Variation: The interquartile range (IQR) for both renovated and non-renovated properties is similar. This suggests that there is a comparable spread of prices within each group, indicating a similar range of prices for properties in both renovation categories.

Outliers: There are outliers present in both groups, represented by individual points located outside the whiskers of the boxes. These outliers signify properties with exceptionally high prices, regardless of their renovation status. This implies that there are properties with elevated prices in both the renovated and non-renovated categories, suggesting that other factors beyond renovation status may influence pricing.

The boxplot analysis indicates that properties that have undergone renovation tend to have higher prices compared to properties without any renovation.

COEFFICIENTS OF THE LINEAR REGRESSION MODEL

Variable	Coefficient
grade	82261.904659
lat	74324.163525
sqft_living	66743.828978
bathrooms	17891.932387
floors	12867.395409
sqft_lot	8034.174725
yr_renovated	2844.173821
sqft_above	1529.221365
long	-1960.092237
bedrooms	-8982.261636
zipcode	-11450.498620
yr_built	-59443.668060

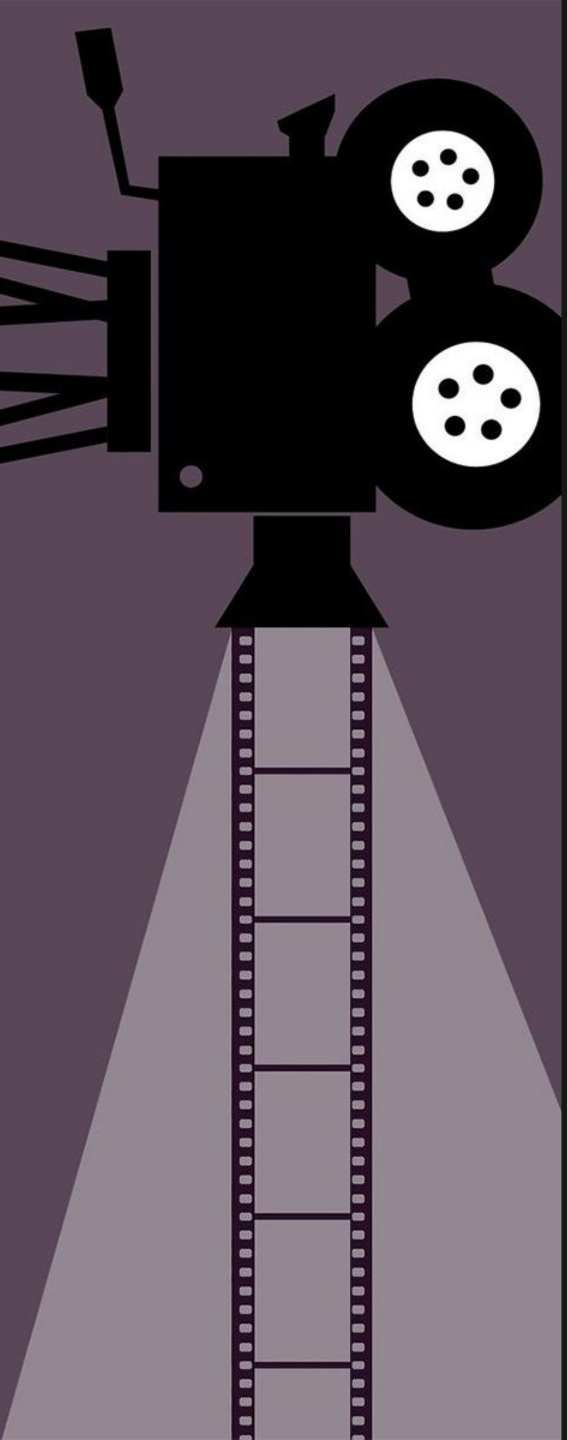


MODELING APPROACH:

The path we followed: I divided the data into training and testing sets, utilizing 80% for training and 20% for testing the model's performance.

Scaling numerical features: To ensure fairness, I standardized the numerical features using StandardScaler to bring them to a comparable scale.

My model of choice: We applied a linear regression model to the training data and employed it to make predictions on the test data.



REGRESSION RESULTS:

Evaluation of our basic linear regression model:

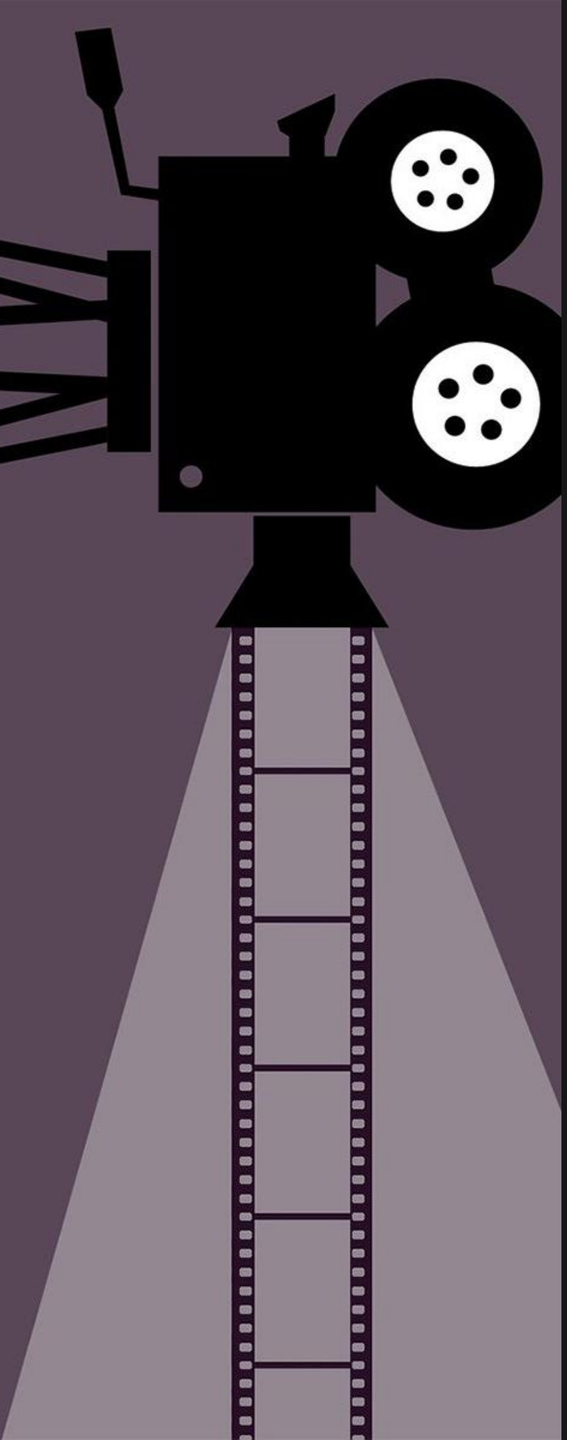
Mean Squared Error (MSE): [MSE value]

R-squared: [R-squared value]

Comparison with other models:

Model	Mean Squared Error (MSE)	R-squared
Basic Linear Regression	13,297,122,013.68	0.6526805215731064
Polynomial Regression	10,855,655,523.80	0.7164513786797331
Cross-Validated Ridge Regression	12,864,564,236.99	0.6667516099567845
Regularized Ridge Regression	13,297,149,685.83	0.652679798779493

Advantages of polynomial regression: The polynomial regression model outperformed the basic linear regression model in terms of MSE and R-squared, indicating superior predictive performance.

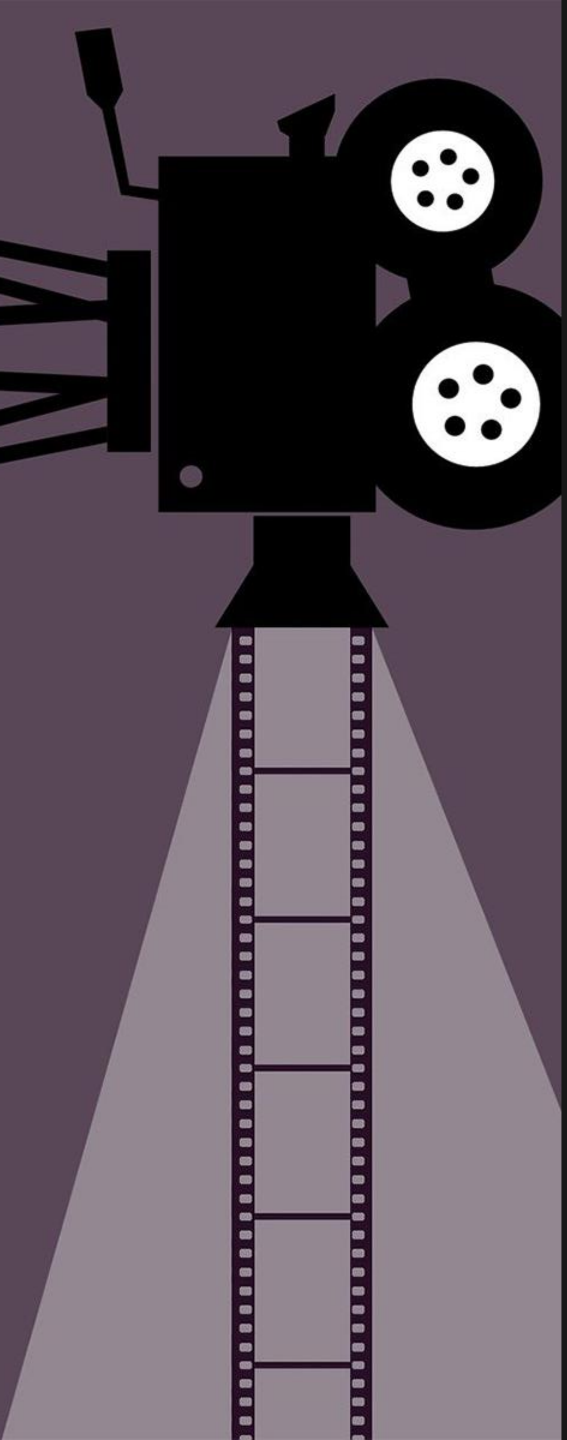


NEXT STEPS:

Continuous refinement: I plan to explore additional feature engineering techniques, investigate interactions between variables, and experiment with alternative algorithms to improve our model's performance.

Expanding data sources: I aim to enrich our analysis by incorporating supplementary datasets, such as neighborhood demographics and local amenities, to capture a more comprehensive understanding of the market.

Ongoing monitoring: I believe in the importance of regularly updating our model with fresh data and continuously monitoring its performance to ensure it remains accurate and reliable.



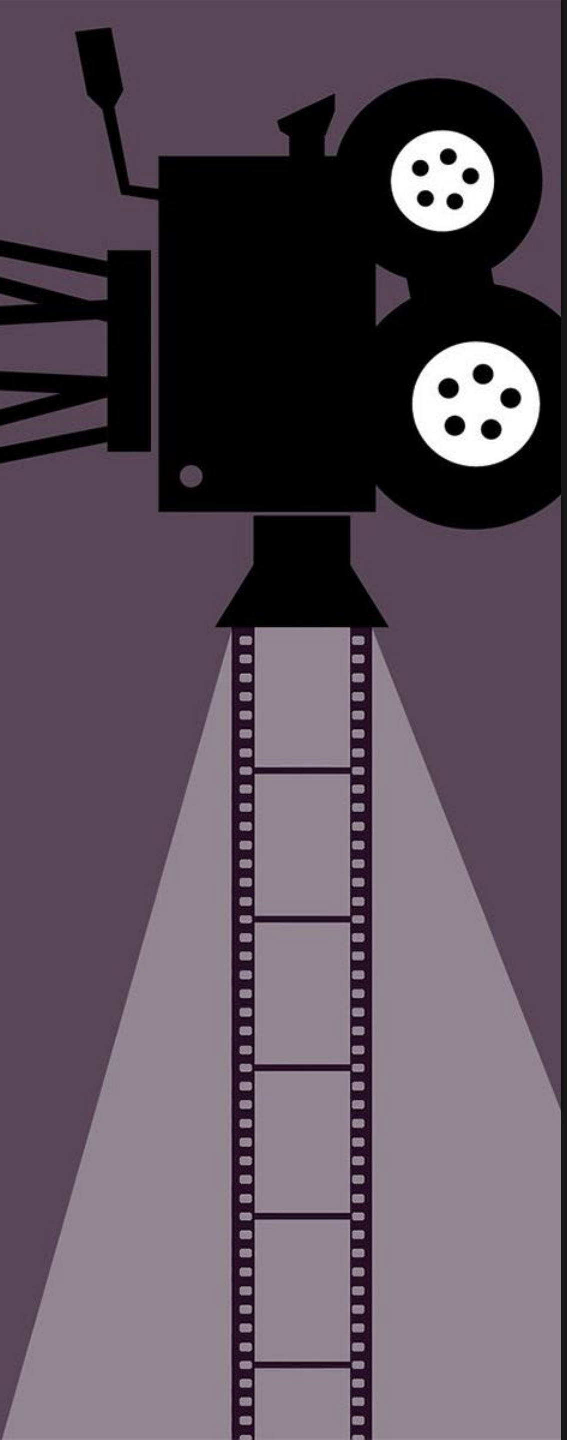
CONCLUSION

In conclusion, this project aimed to predict house prices based on various features and provide insights for buyers and sellers in the real estate market. By analyzing the dataset, exploring the data, and building a regression model, valuable insights were gained into the factors influencing house prices.

The findings indicate that the property grade, location, living space, number of bathrooms, and number of floors have the most significant impact on house prices. Other factors such as lot size, renovation status, square footage above ground, number of bedrooms, specific zipcode, and property age also contribute to price variations.

These recommendations are meant to guide buyers and sellers in making informed decisions in the real estate market. It is important to consider these factors alongside personal preferences, market conditions, and seek professional advice when navigating the dynamic real estate market.

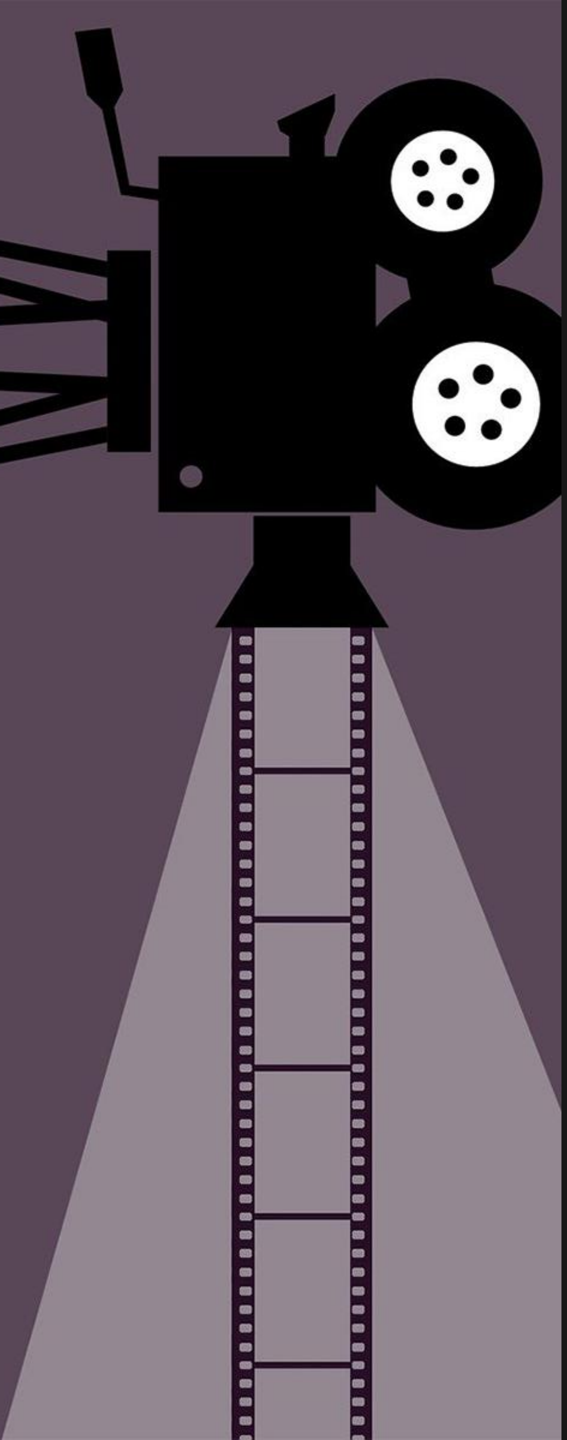
Overall, this project provides valuable information and insights that can assist in understanding the factors influencing house prices, helping individuals make more informed choices in their real estate endeavors.



RECOMMENDATIONS:

Based on the analysis and insights gained from the modeling results, here are some recommendations for buyers and sellers in the real estate market:

1. Focus on property grade: As a seller, consider investing in renovations and upgrades to improve the grade of your property. A higher grade can attract more buyers and potentially lead to higher selling prices. As a buyer, prioritize properties with a good grade, as they are likely to offer better quality and higher value.
2. Consider location: Pay attention to the location of the property. Desirable locations tend to command higher prices. Buyers should prioritize properties in popular and convenient areas, while sellers can highlight the advantages of their property's location to attract potential buyers.
3. Invest in living space: The size of the living area has a significant impact on the price of a house. Buyers looking for more space and higher value should consider properties with larger living areas. If you're a seller, consider adding extensions or maximizing the existing living space to increase the value of your property.
4. Bathrooms add value: The number of bathrooms in a property also influences its price. Buyers should consider properties with an adequate number of bathrooms to meet their needs and potentially increase the property's value. Sellers can consider adding or renovating bathrooms to make their property more appealing to buyers.



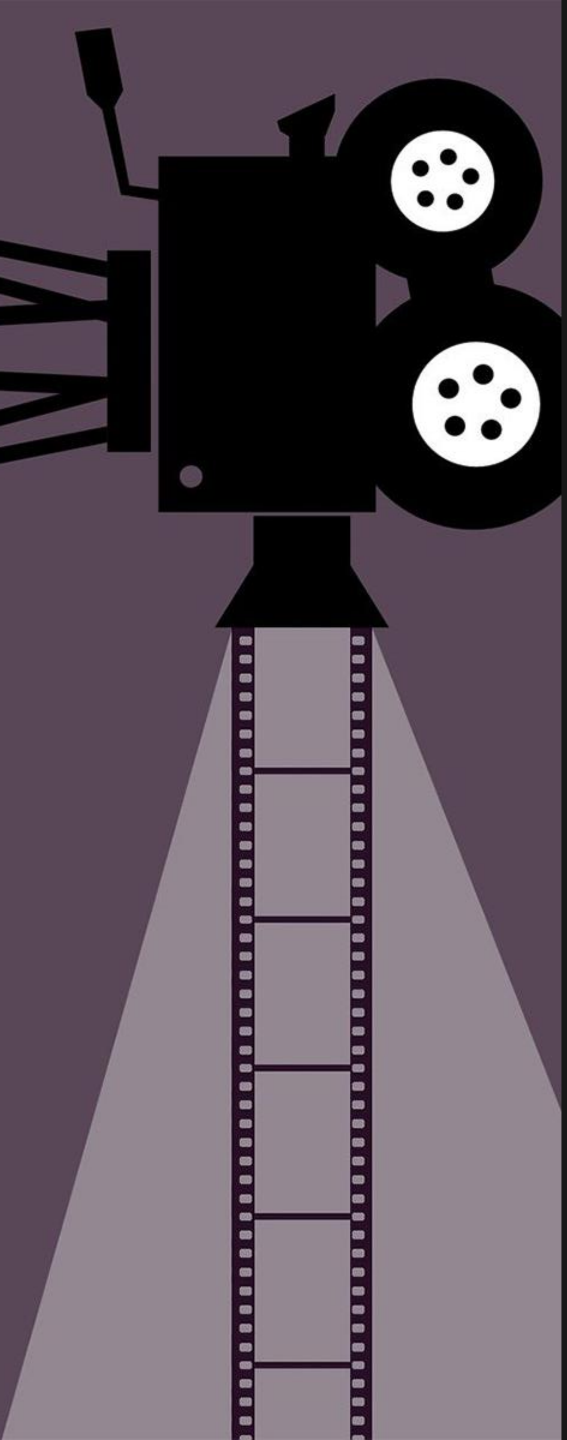
RECOMMENDATIONS:

5. Consider additional floors: While the number of floors alone may not strongly influence the price, properties with multiple floors can offer more space and potentially higher value. Buyers seeking more space or unique architectural features may explore properties with multiple floors. Sellers can emphasize the advantages of additional levels to attract interested buyers.

6. Lot size and renovation: Although the impact is relatively smaller, factors such as lot size and renovation can affect the price. Buyers should consider their preferences regarding lot size and weigh them against other influencing factors. Sellers can highlight any renovations or improvements done to their property to justify the asking price.

7. Bedrooms and zipcode: The number of bedrooms and the specific zipcode can have a negative impact on house prices. Buyers looking for more affordable options may consider properties with fewer bedrooms or explore properties in different zipcodes. Sellers should be aware of the number of bedrooms and consider pricing strategies to align with the market.

8. Property age: The year a house was built can affect its price. Buyers who prefer newer properties should focus on recently constructed or renovated houses. Sellers of older properties can highlight unique features or historical value to compensate for the age factor.



THANK YOU:

Expressing gratitude: I sincerely thank you for your time and attention today. It has been a pleasure sharing our analysis with you.

Contact information: Should you have any further inquiries or wish to collaborate on related projects, please feel free to reach out to me.