House Sales in King County, USA

Analysis and Predictions

•••

Orlando Marin

Topics of Discussion

1. Introduction to the Dataset
2. Data Exploration
3. Modeling Overview and Results
4. Insights and Comparison
5. Conclusion and Takeaways

Introduction to the Dataset

Introduction to the Dataset

- The “House Sales in King County, USA” dataset contains house sale prices for King County, which includes Seattle, WA. It includes homes sold between May 2014 and May 2015.
- Why did I choose this dataset?
    - Buying and Selling Property - Understand the factors that influence home prices, helping students make informed decisions when buying or selling a property.
    - Renovating Property - Identify which home improvements provide the best return on investment, guiding students on how to increase property value for resale.
    - Inheritance or Estate Planning - Students may inherit property and become involved in decisions related to selling a relative’s home. Understanding home values can help make decisions regarding selling or maintaining property.

Kaggle Dataset

Data Exploration

Data Cleaning, Prep, and Feature Engineering

- Ensured that the "date" column was converted from a string to datetime
- Dropped the "id" column since it is a unique identifier, and there were no duplicates
- Treated the "zip_code" column as categorical instead of an integer
    - Ultimately removed the "zip_code" column since to avoid redundancy
    - The dataset had latitude and longitude data
    - There were 70 different zip codes
- Removed rows of homes that had either 0 bedrooms or 0 bathrooms
- Replaced the "yr_renovated" column values of "0" with the corresponding value from the "yr_built" column for homes that were never renovated
- Created a binary "has_basement" column
    - 0 means the home doesn't have a basement
    - 1 means the home has a basement
- Removed the "sqft_above" and "sqft_basement" columns to avoid redundancy with "sqft_living"
    - The sum of "sqft_above" and "sqft_basement" equals the "sqft_living" column
    - I created the "has_basement" column
- Removed rows with outliers for "price", "sqft_lot", and an extreme "bedrooms" outlier (33 bedrooms)

Google Colab

Dataset Features

Original Features (21 columns, 21613 rows)

- id
- date (string)
- price
- bedrooms
- bathrooms
- sqft_living
- sqft_lot
- floors
- waterfront
- view
- condition
- grade
- sqft_above
- sqft_basement
- yr_built
- yr_renovated (many rows with value of “0” for homes never renovated)
- zip_code
- lat
- long
- sqft_living15
- sqft_lot15

Features after Modifications (18 columns, 18229 rows)

Features Removed: id, sqft_above, sqft_basement, zip_code
Features Updated: date, price, bedrooms, sqft_lot, yr_renovated
Features Created: has_basement

- date (datetime)
- price (removed outliers)
- bedrooms (removed an extreme outlier)
- bathrooms
- sqft_living
- sqft_lot (removed outliers)
- floors
- waterfront
- view
- condition
- grade
- yr_built
- yr_renovated (homes that had “0” were replaced with yr_built)
- lat
- long
- sqft_living15
- sqft_lot15
- has_basement

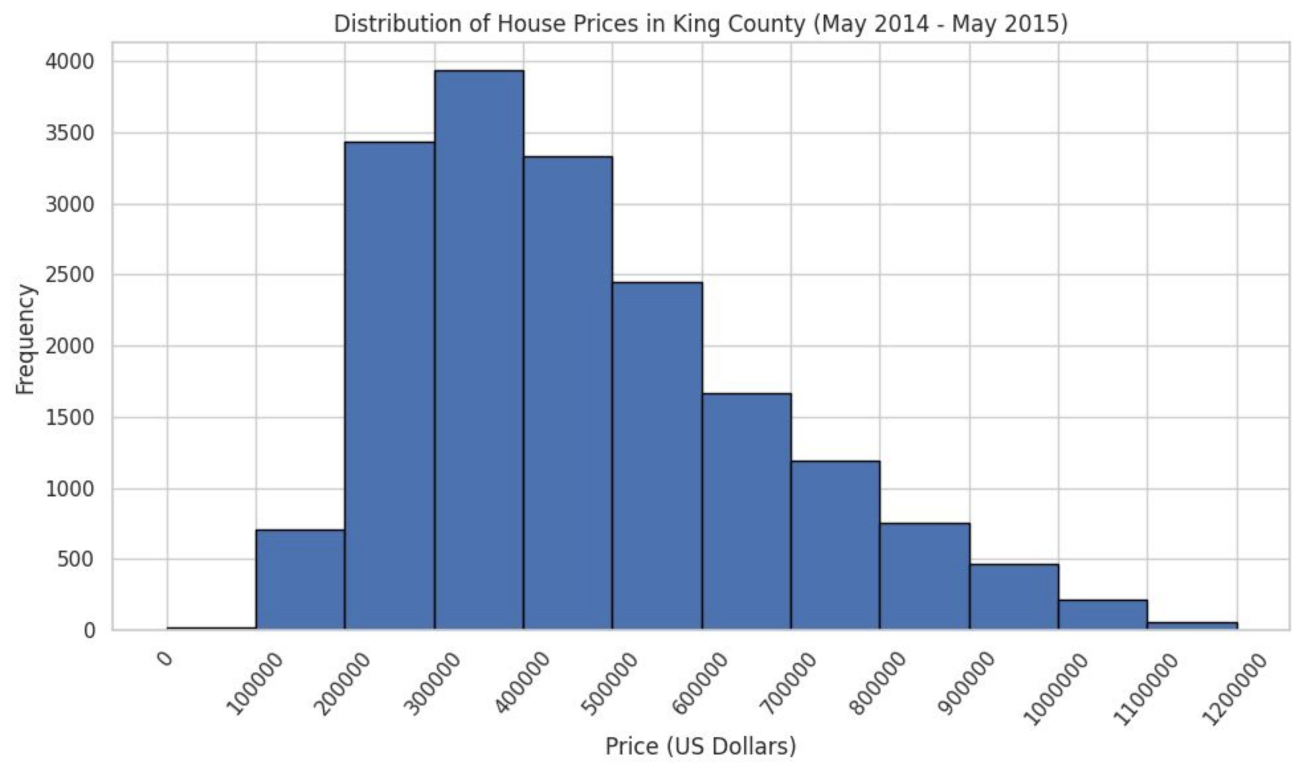
[Google Colab]()

Summary statistics of key features after data cleaning

	price	bedrooms	bathrooms	sqft_living	sqft_lot	grade	yr_built
Mean	$467,147	3.31	2.02	1910.65	7214.46	7.47	1970
Standard Deviation	$204,677	0.89	0.71	727.76	3461.94	0.98	29.92
Minimum	$78,000	1.00	0.50	370.00	520.00	3.00	1900
50%	$426,000	3.00	2.00	1800.00	7189.00	7.00	1972
Maximum	$1,127,000	11.00	7.50	7350.00	18295.00	12.00	2015

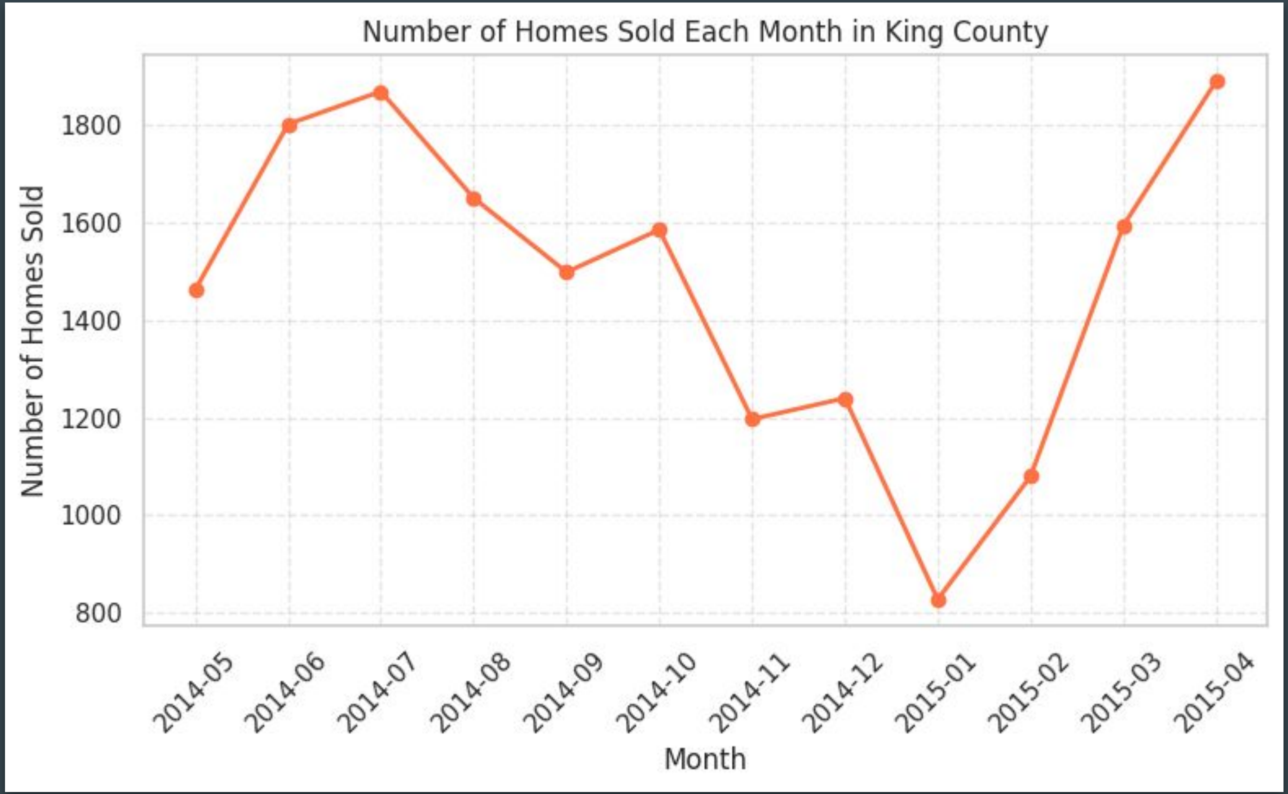
Google Colab

Most home prices in King County are between $200K and $500K



[Google Colab]

Most homes in King County are sold in spring and summer months

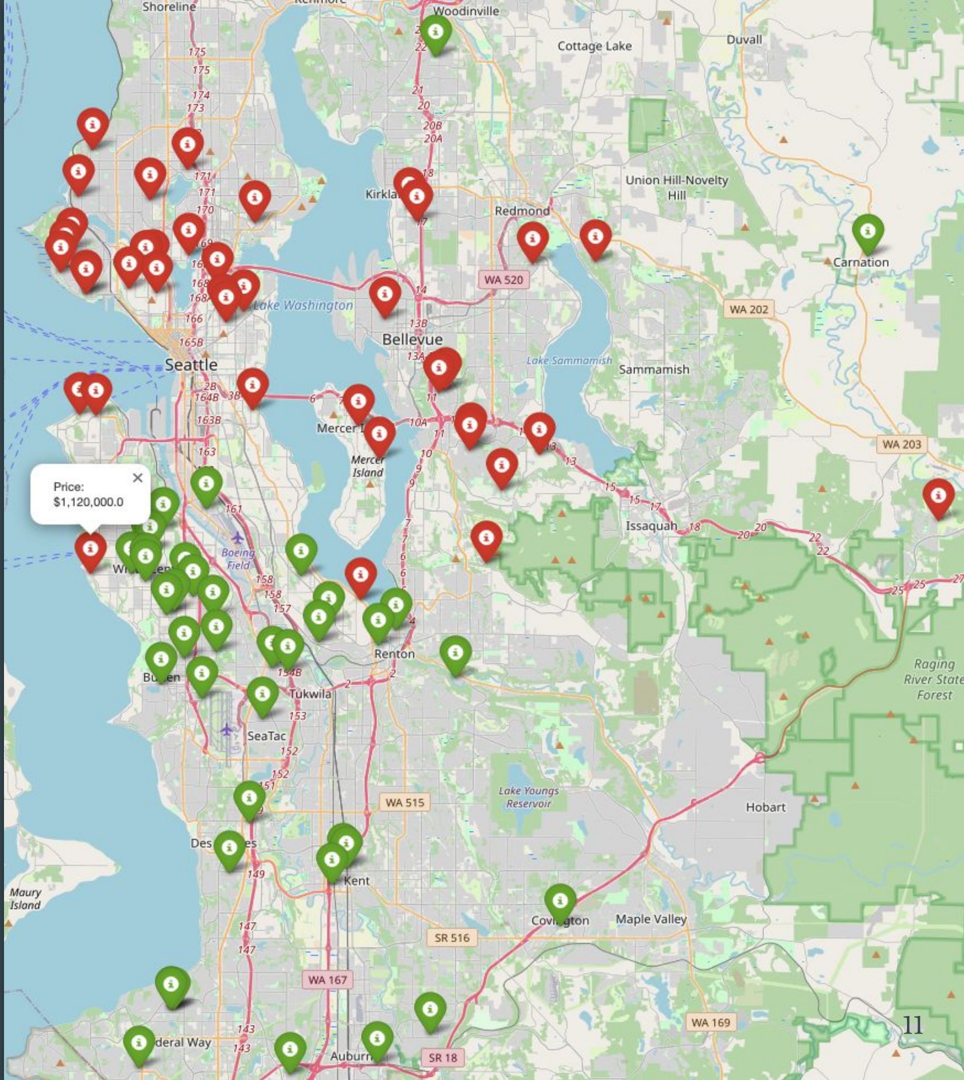


Google Colab

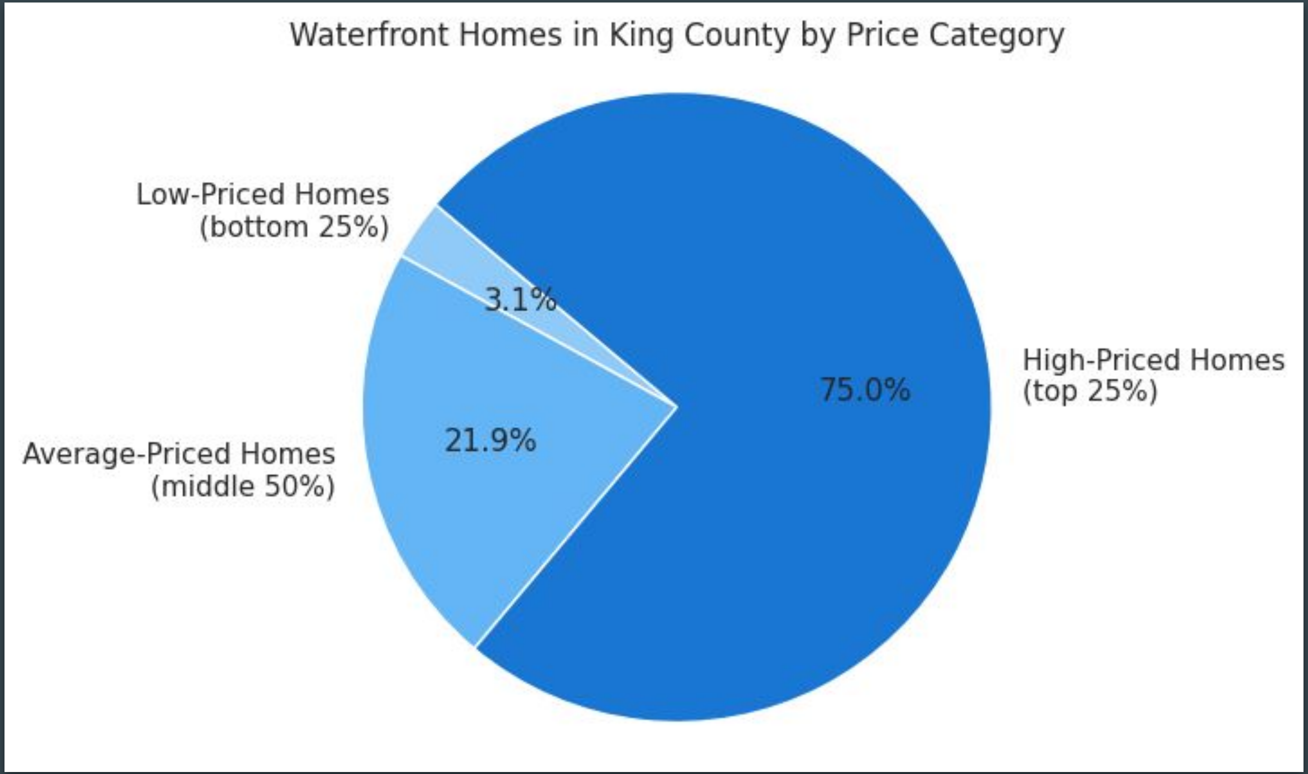
Locations of the most and least expensive homes in King County

- The 40 most expensive homes in King County have red markers
- The 40 least expensive homes in King County have green markers
- The highest priced homes in King County tend to be further north
    - This indicates that latitude may be a more important factor when predicting home price than longitude
- The highest priced homes also tend to be near Seattle and Bellevue, two major cities, and/or near the water

[Google Colab]

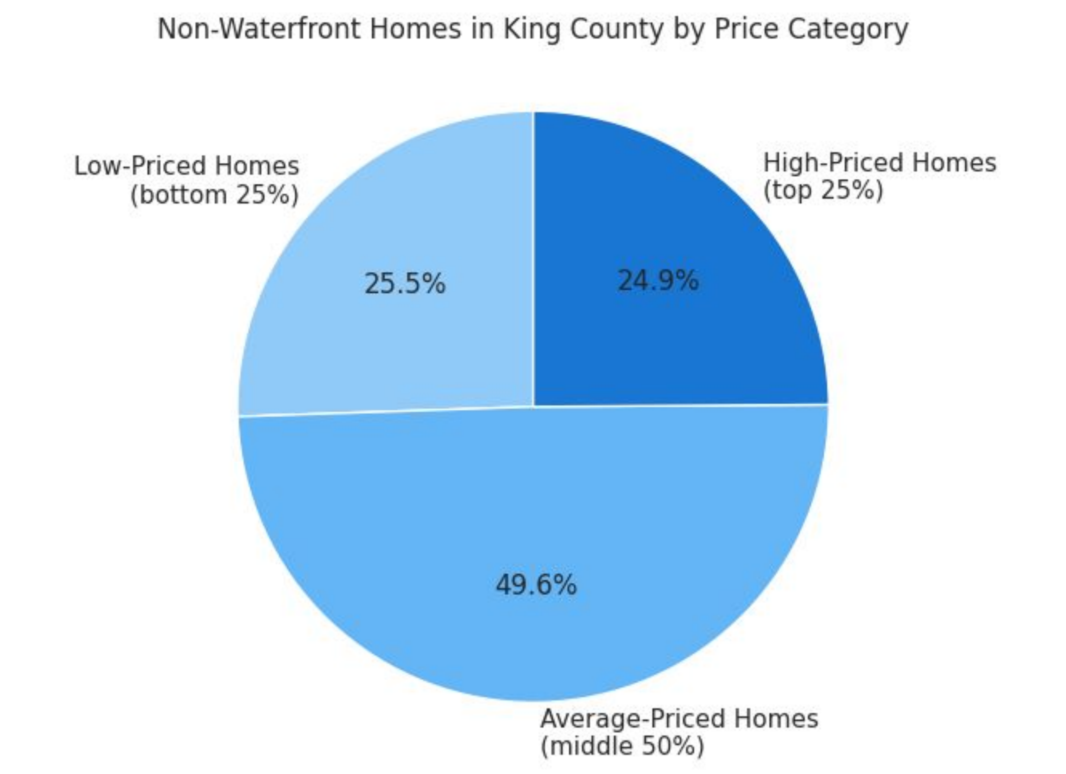


The vast majority of waterfront homes are high-priced



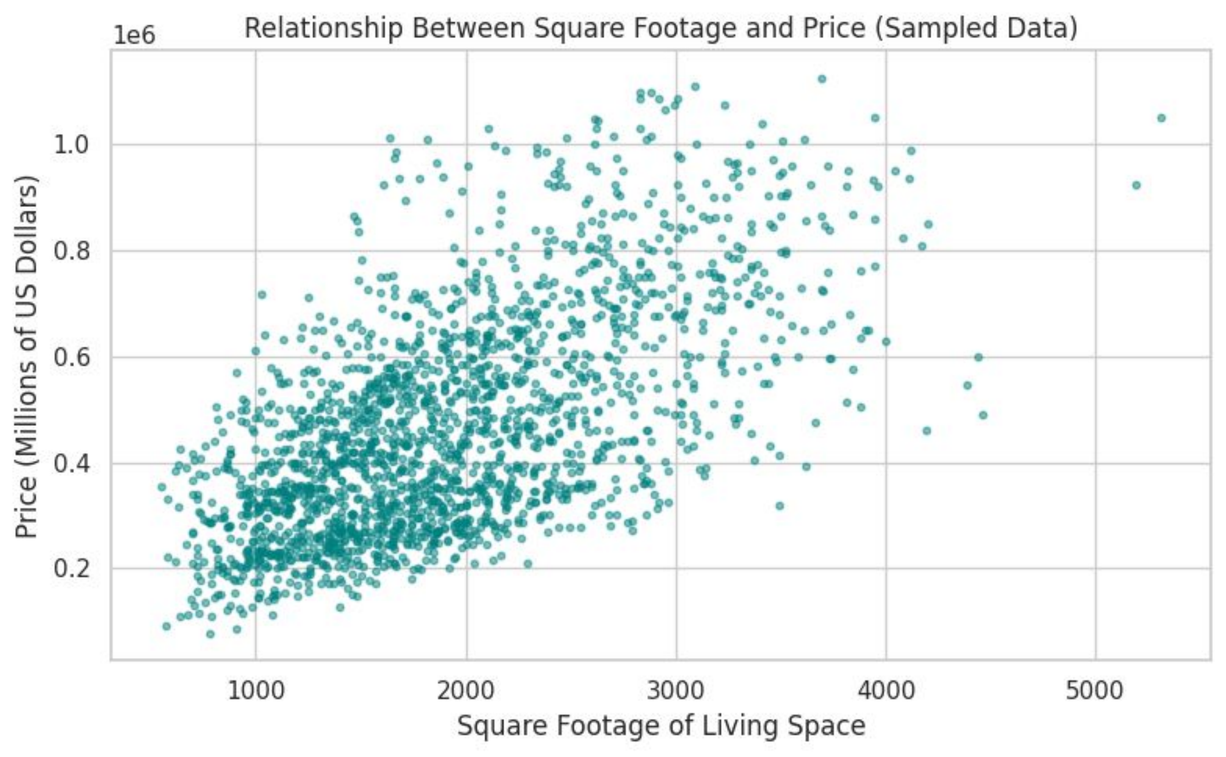
[Google Colab]

Non-waterfront homes have a balanced distribution by price range



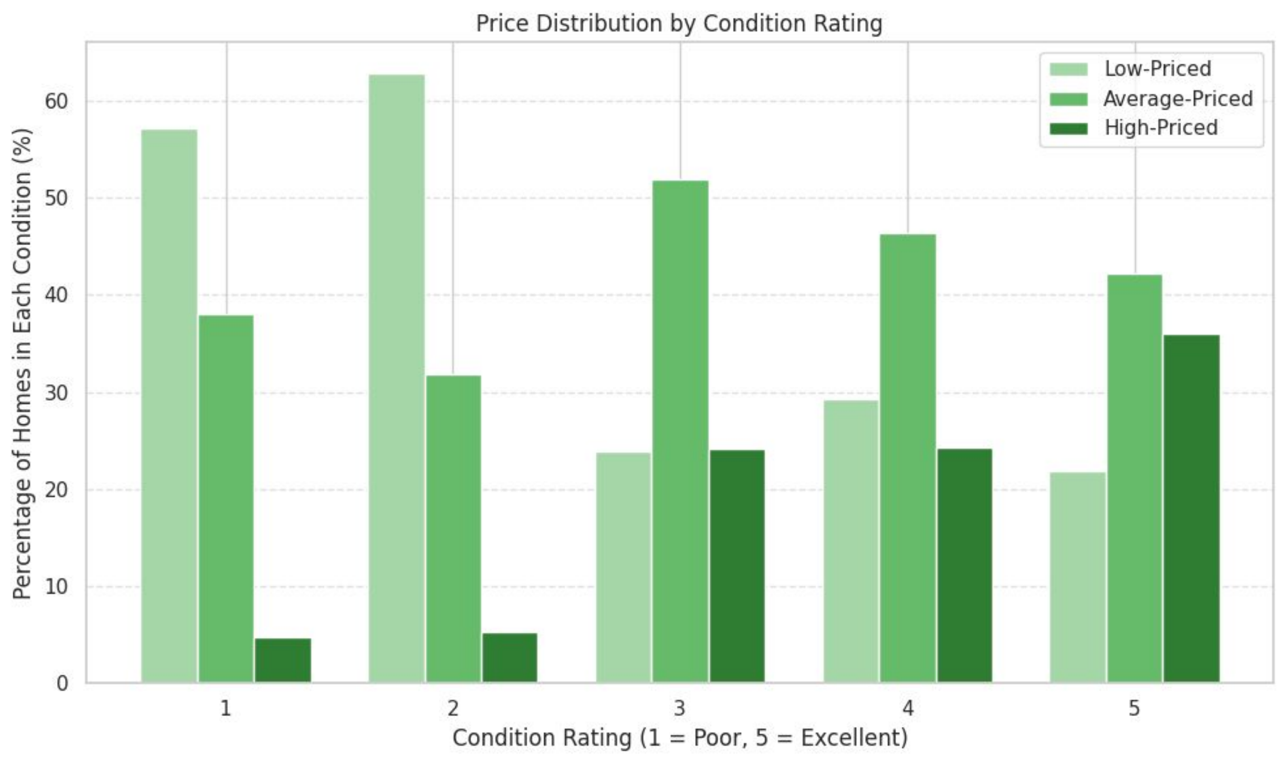
[Google Colab]

There is a linear relationship between sqft_living and price



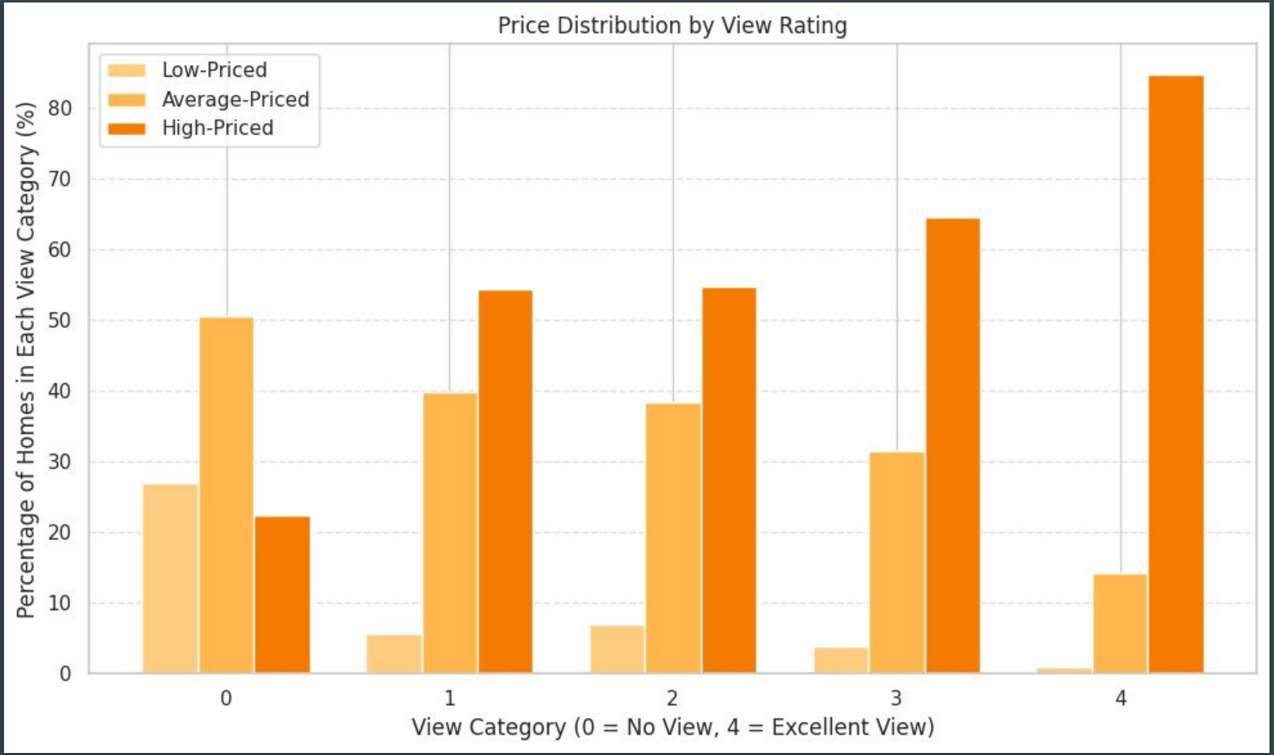
[Google Colab]

As condition increases, so does the proportion of high-priced homes



[Google Colab]

As the view improves, so does the proportion of high-priced homes



[Google Colab]

Modeling Overview and Results

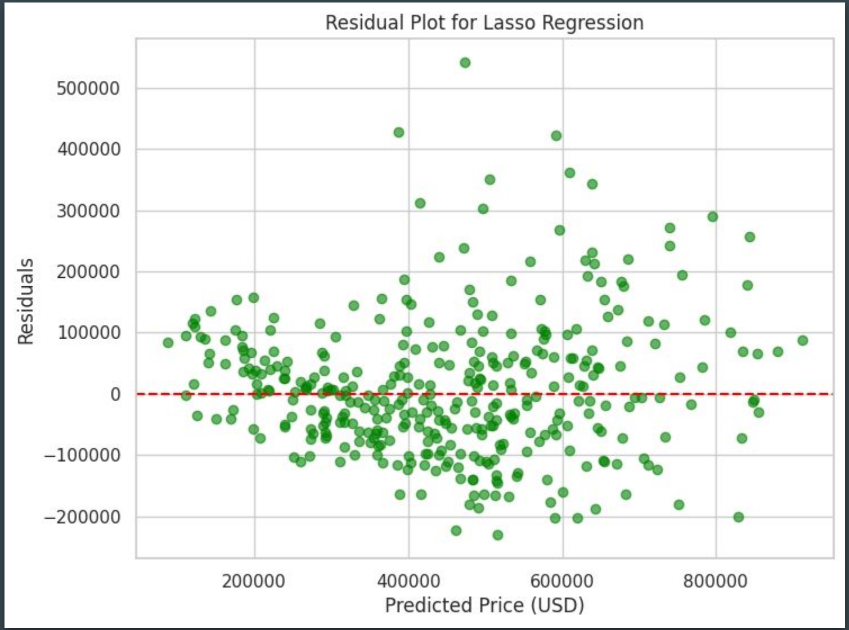
Lasso Regression Model

Purpose - apply Lasso regression to predict house prices in King County and compare its performance to KNN regression, leveraging regularization to handle overfitting and feature selection

Method - implemented a Lasso regression model with alpha set to 1.0. Data was split into 80/20 training/testing sets, and numerical features were standardized before model fitting

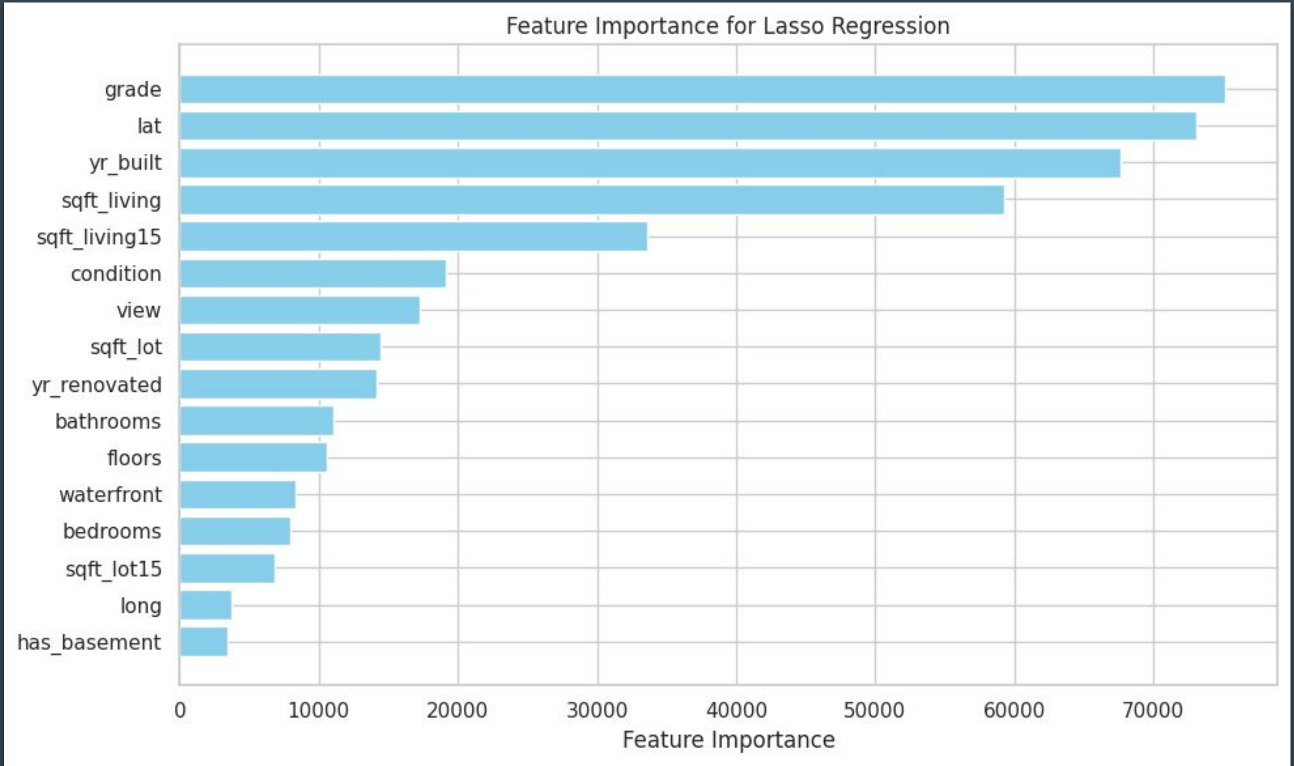
Results

- R^2 Score: 0.7011
- Root Mean Squared Error: $112,313
- Mean Price: $469,133



Google Colab

Feature Importance - Lasso Regression



[Google Colab]

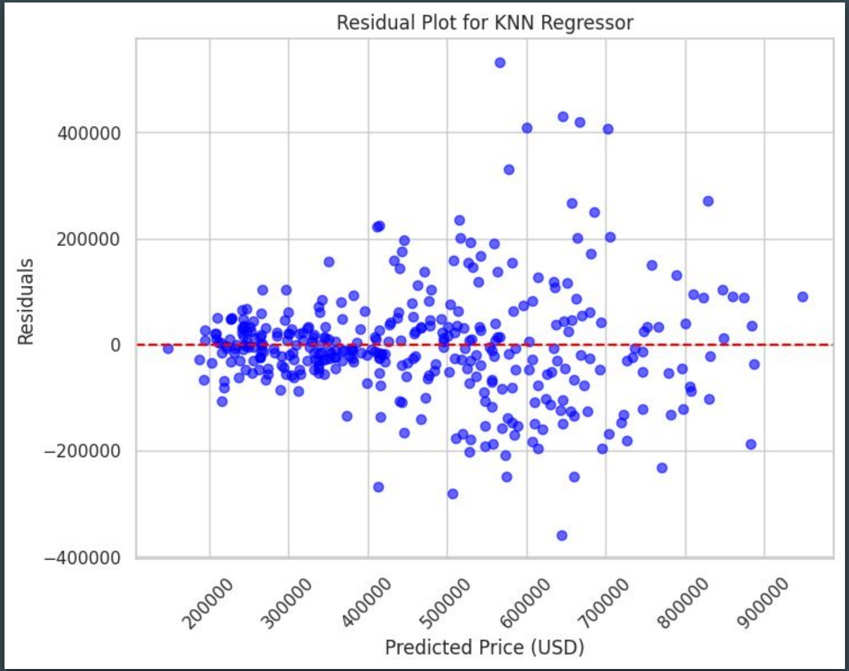
KNN Regressor Model

Purpose - Apply KNN regression to predict house prices in King County and compare its performance to Lasso regression, using non-linear relationships between features.

Method - implemented KNN with n_neighbors = 5. Data was split into 80/20 training/testing sets, and numerical features were standardized before model fitting.

Results

- R^2 Score: 0.7721
- Root Mean Squared Error: $98,078
- Mean Price: $469,133



Google Colab

Insights and Comparison

Which model performed best, and why?

- Best performing model: KNN Regression
  - KNN performed better with an R^2 of 0.7721, compared to Lasso's 0.7011. This indicates that KNN explains more of the variance in King County house prices.
  - KNN also had a lower Root Mean Squared Error of $98,078 compared to Lasso's $112,313, indicating that KNN's predictions were closer to the actual values.
- Lasso Regression
  - Helped identify which features mattered most, such as grade, sqft_living, and latitude, which gives useful insight into what drives house prices in King County.
  - Lasso uses regularization to reduce the influence of less important features, helping to avoid overfitting and making the model easier to interpret.
- Why KNN performed better?
  - KNN can model non-linear patterns in the data. It can detect more complex relationships between the features and house prices that a linear model like Lasso might miss.

[Google Colab]

Conclusion and Takeaways

Conclusion and Takeaways

Limiting Factors

- Important details like school district, crime rate, etc. were not included in the dataset, even though they can significantly impact house prices.
- The models were trained only on King County data, so the results may not apply well to other housing markets.

Suggestions for Future Study

- Test another model like Random Forest to potentially improve accuracy.
- Add features like crime rate, school ratings, or walkability scores to build a more complete model of what influences house prices

Recommendations to Students

- When buying a home, prioritize homes in great condition with larger living space and desirable location features (view, waterfront) to maximize long-term value.
- When selling and renovating a home, invest in upgrades that improve the home's condition and living area to have the biggest impact on price.

Thank you!