Location, location, location:

A multi-linear regression model for home price prediction

Kevin Giroux

September 2020

Problem statement

This analysis was performed for a real estate investment fund, with the following goals:

- **Primary goal**: Help the fund to accurately price homes in their inventory for future sale

- **Secondary goal**: Provide insight into how various factors affect the predicted sale price of home, with a particular focus on the 'zipcode' variable

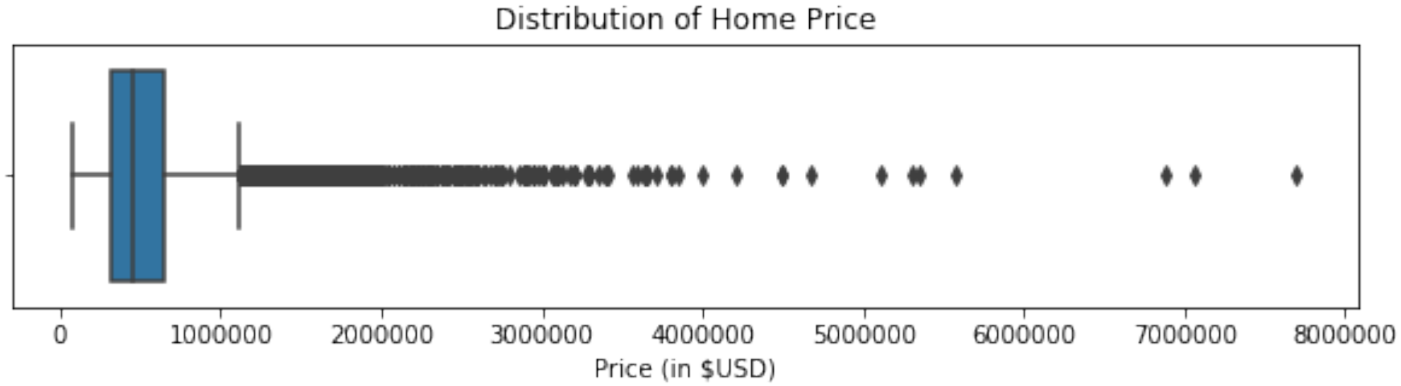
Dataset overview

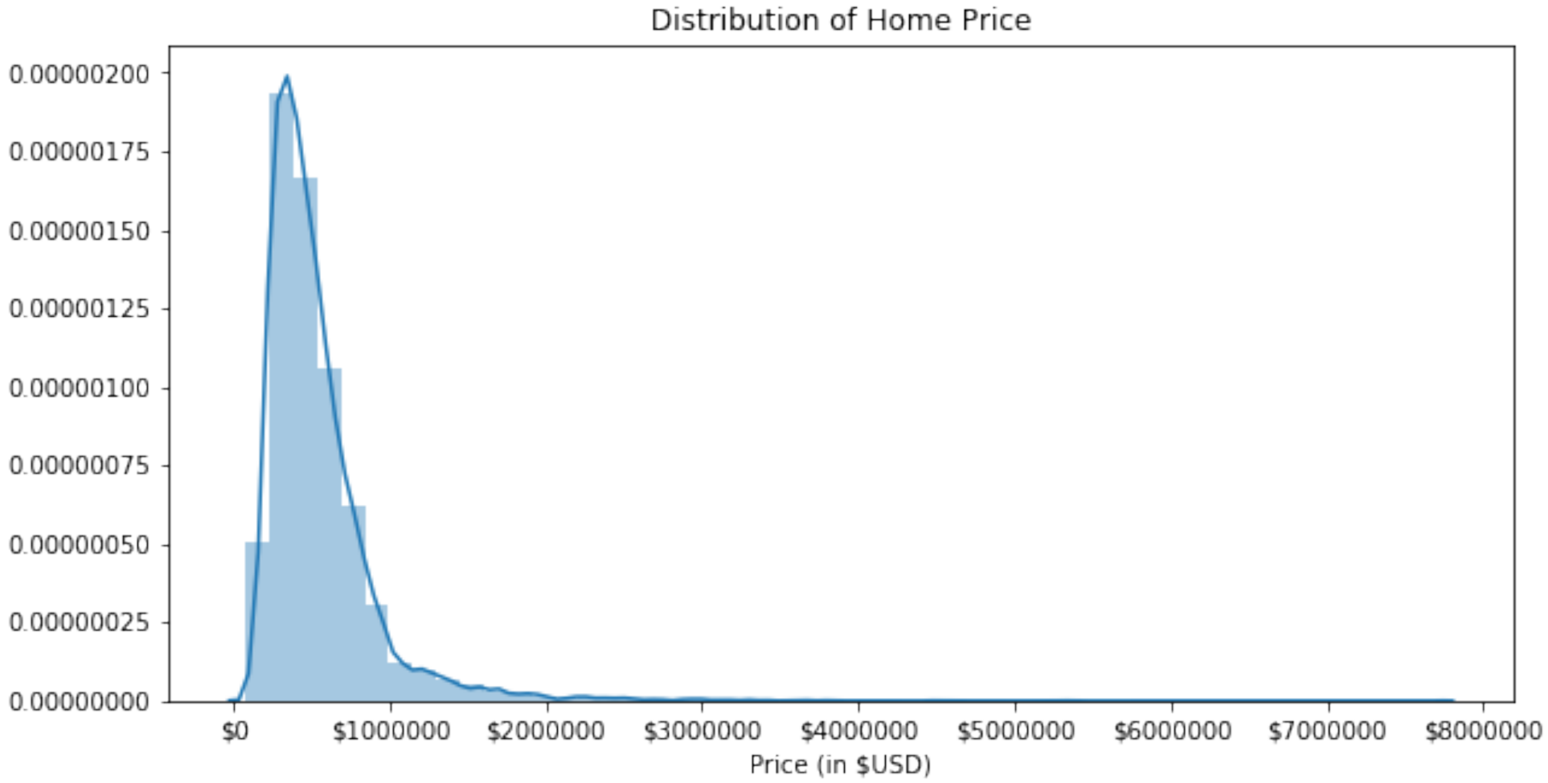
- For this analysis, I used the King County House Sales dataset, which details the many physical attributes and the corresponding sale prices of a sample of approximately 21k homes, all located in the Seattle, Washington area.

- The following features were included in the data, with additional detail as necessary

    - Sale dates
    - Sale price
    - Bedrooms (count)
    - Bathrooms (count)
    - Living sqft
    - Lot sqft
    - Floors (count)
    - Waterfront (binary variable representing whether or not the home is on the water)
    - View (count of how many times a home has been viewed)
    - Condition (numerical rating of home condition)
    - Grade (numerical rating of home condition)
    - Above ground sqft
    - Basement sqft
    - Year built
    - Year renovated
    - Zipcode
    - Latitude + Longitude (coordinates)
    - Neighbors (for each home, the average square-footage of both the nearest 15 homes AND their respective lots)

Dataset overview





HOME PRICE:

Sample size: 21,597 homes

Mean price: $540,296

Median: $450,000

Min: $78,000

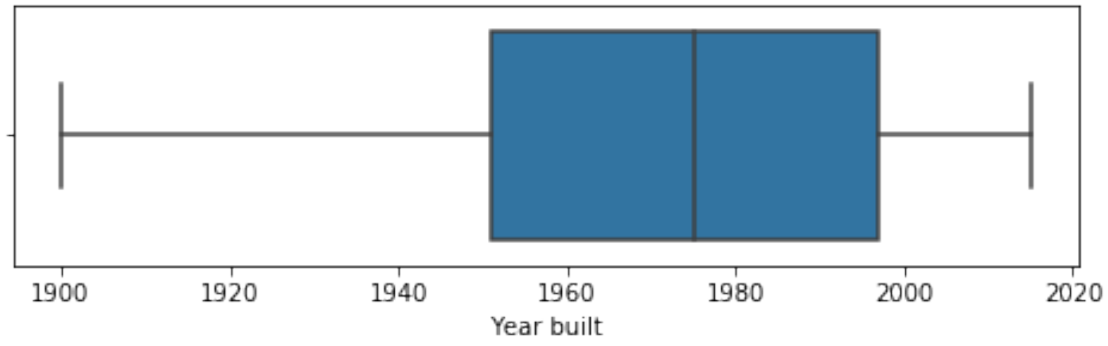
Max: $7,700,000

Conclusion: Outlier removal necessary prior to building a predictive model

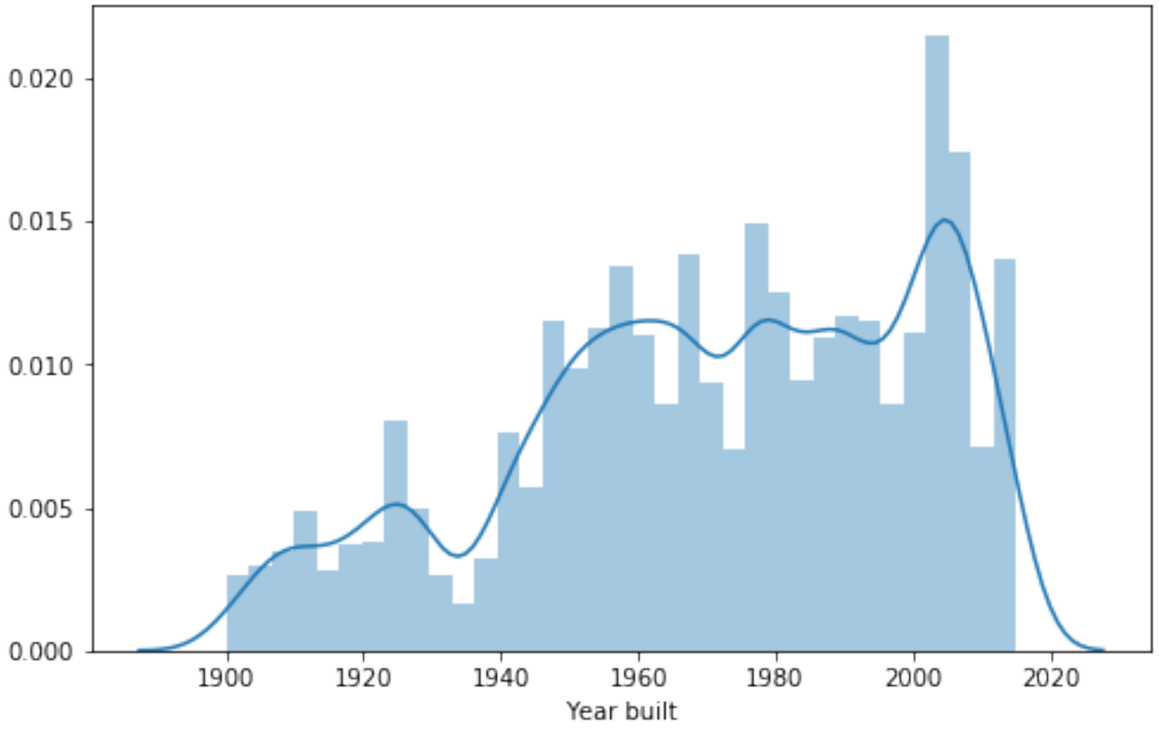
Dataset overview

YEAR BUILT:

Distribution of Year Built

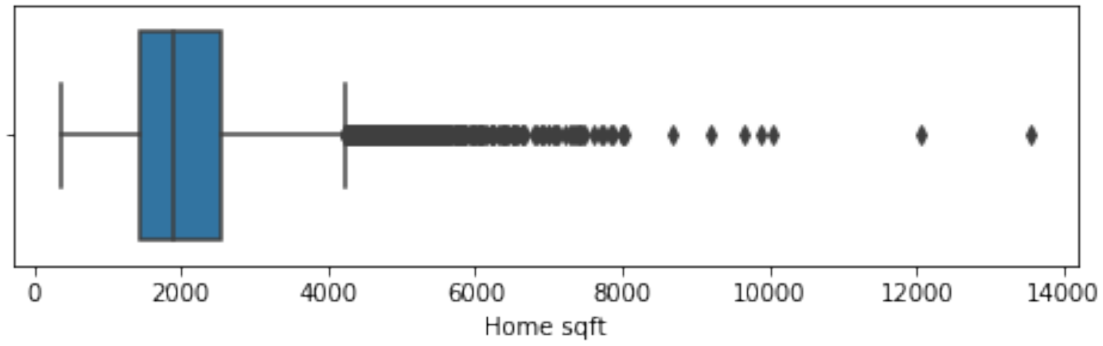


Distribution of Year Built

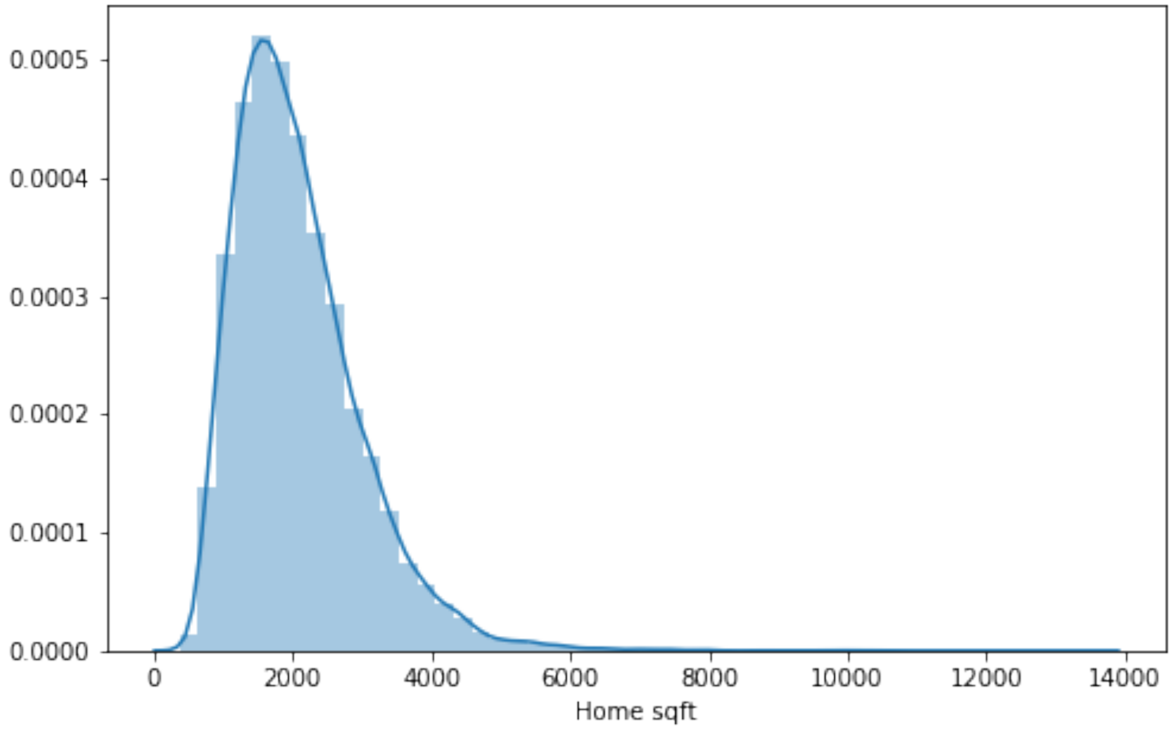


HOME SQFT:

Distribution of Living Area Square-footage

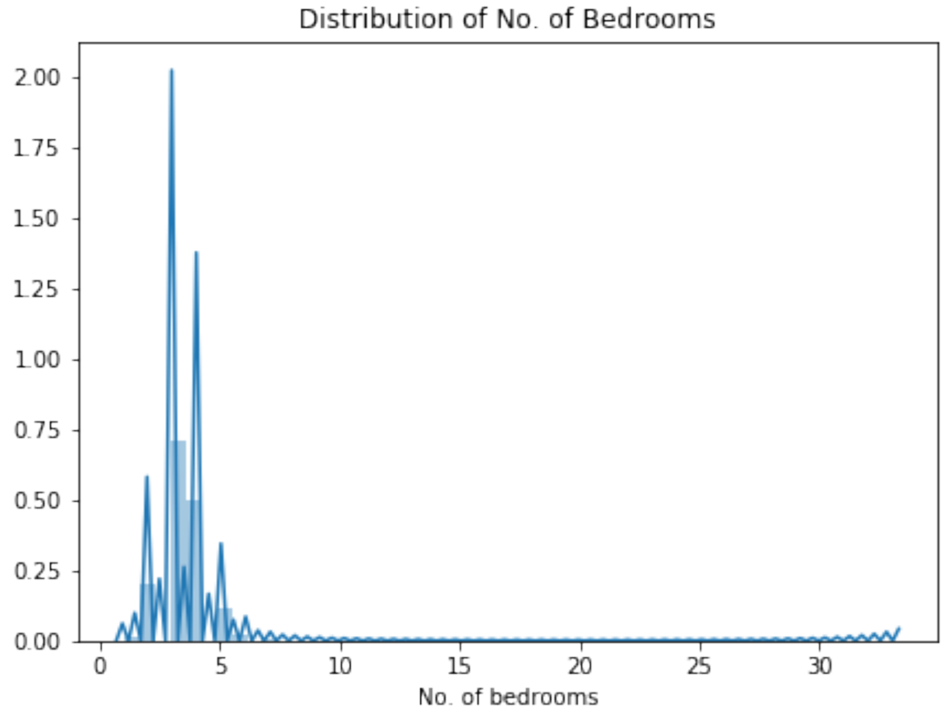


Distribution of Living Area Square-footage

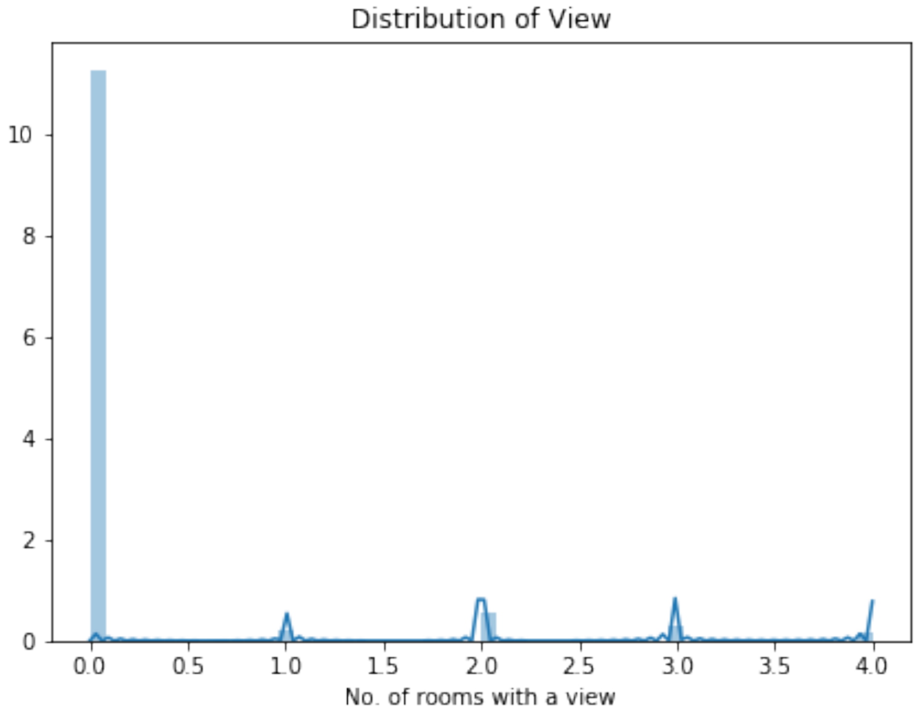


Dataset overview

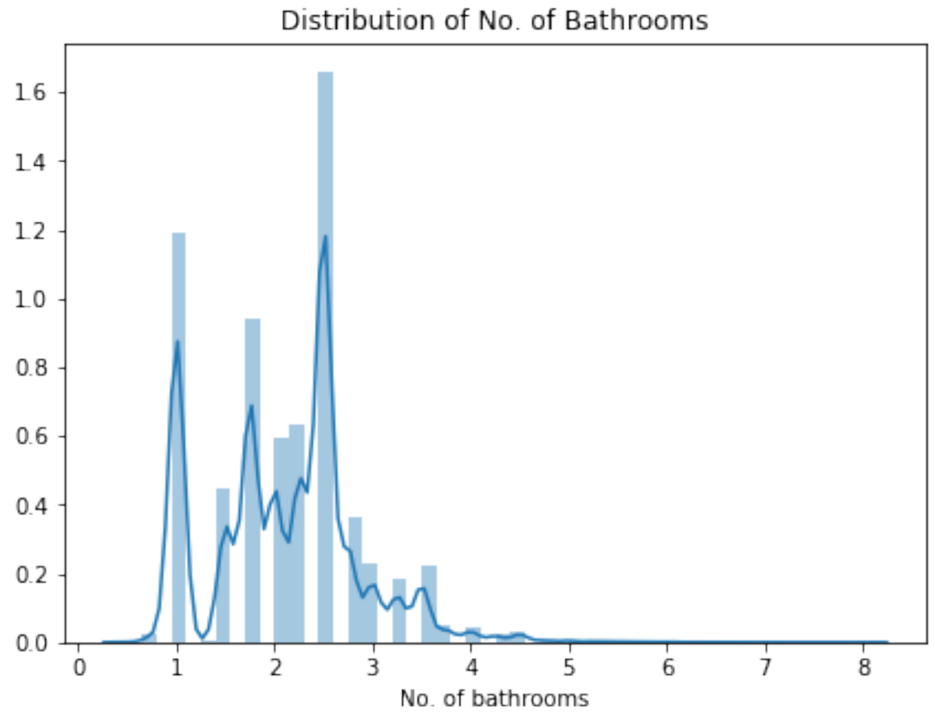
BEDROOMS:



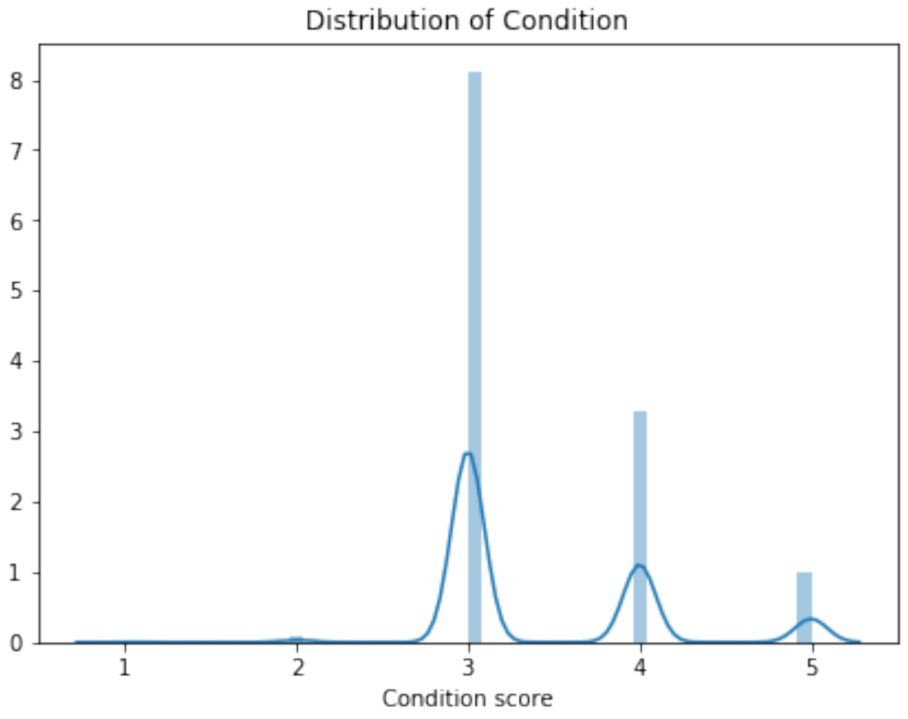
NO. OF TIMES HOME HAS BEEN VIEWED:



BATHROOMS:



HOME CONDITION:



ZIP CODES!

There are 70 different zip codes represented in the sample data; can we use them to assist in price prediction?

Methodology & Limitations

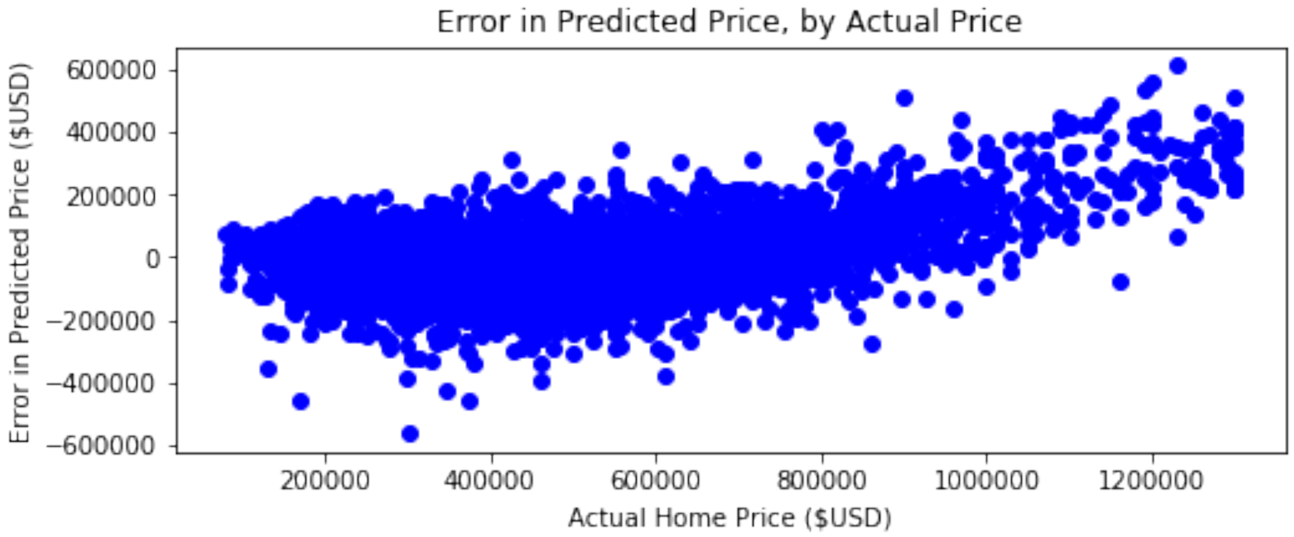
- Analysis of each feature in the sample data, and removal of outliers
- As a result, our model was trained to best predict prices of homes with the following characteristics:
    - Home price <= $1.3mm
    - Bedrooms <= 6
    - Living sqft <= 4500
    - Lot sqft <= 17,500
- Trained a multi-linear regression model and then tested it with new data to confirm model's predictive power
- Used the resulting model to answer the following questions:

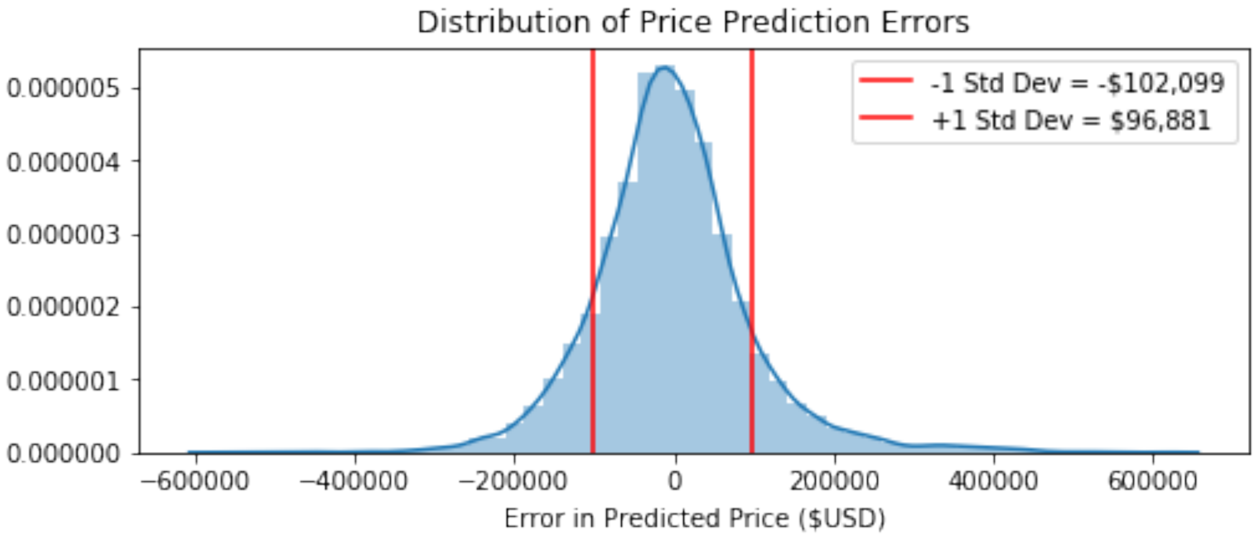
Questions for analysis:

- How accurately can the model predict the price of a home?
- Is ‘zipcode' useful as a predictor of home price?
- Which other variables have the strongest impact on predicted home price?

RESULTS

1. How accurately can the model predict home price?


Error in Predicted Price, by Actual Price


Distribution of Price Prediction Errors

- Model Adj. R-squared value: .797
- In other words, our model is able to explain 79.7% of the observed variation in price
- 68% of the time the model’s predicted price is within $100k of the homes actual value

Recommendation:

- This model is useful for predicting home price with approximately $100k of the actual value of a home; as such, it MUST be used in conjunction with our collective professional expertise and experiences for the precise valuation of our inventory.

2. Is ‘zipcode’ useful as a predictor of home price?

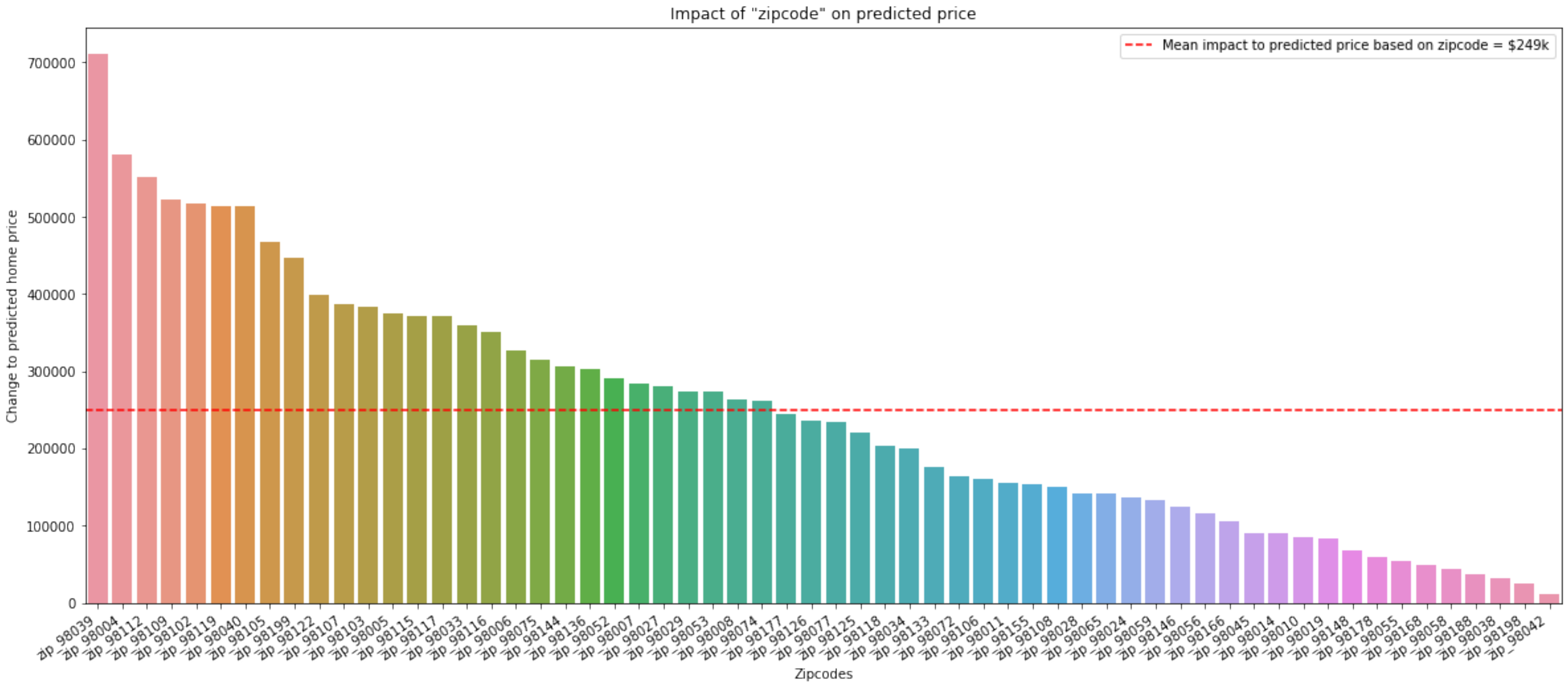
- YES!

- The final model contains variables representing 60 different zip codes and prescribes a unique value to the predicted price of a home based on which zip code that home is located in

Recommendations:

- Bump up the prices of any of our listed inventory in the most expensive zipcodes;
- Institute caps on the bids we make for investments in cheaper zip codes

2. Is ‘zipcode’ useful as a predictor of home price?

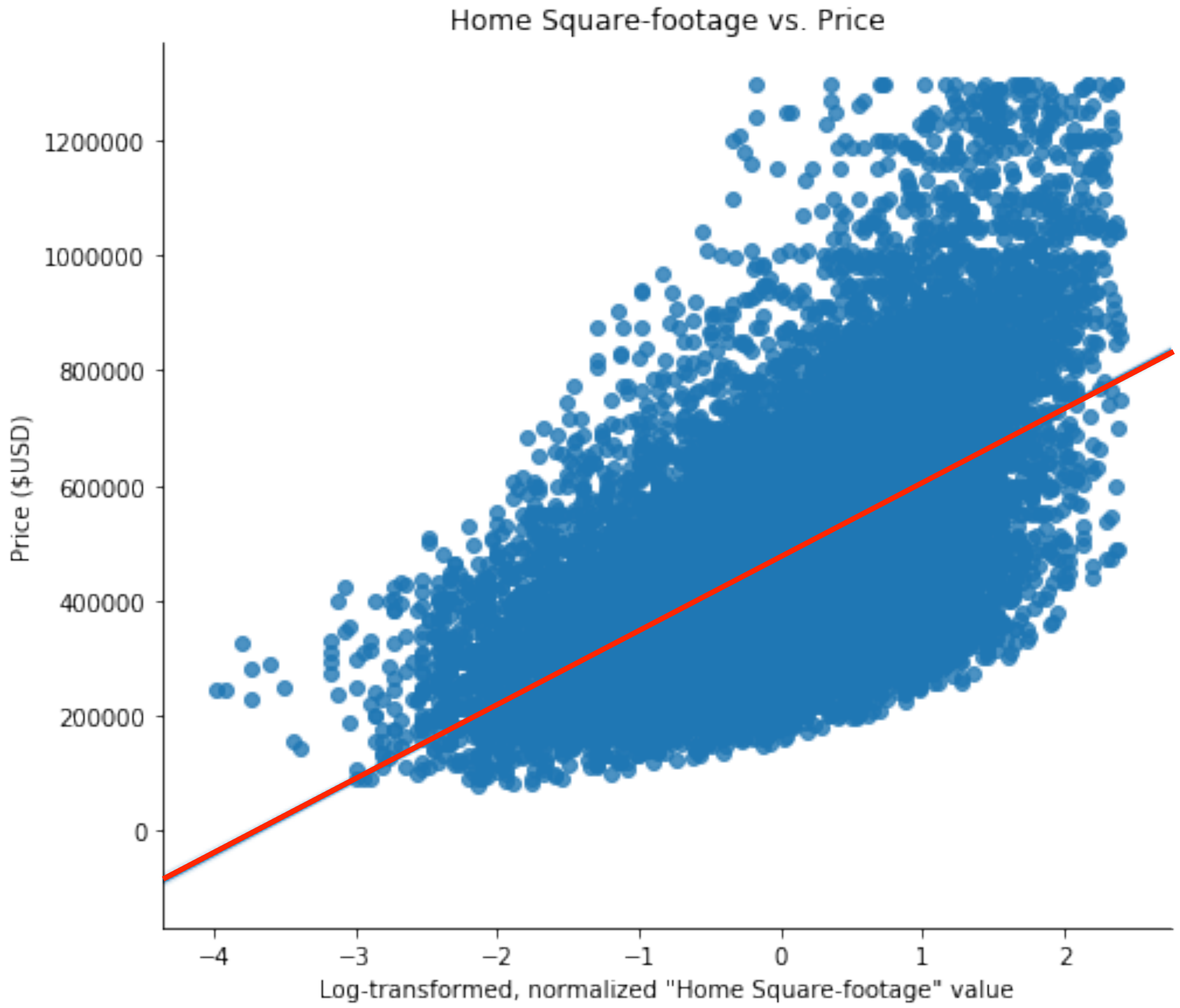


Recommendations:

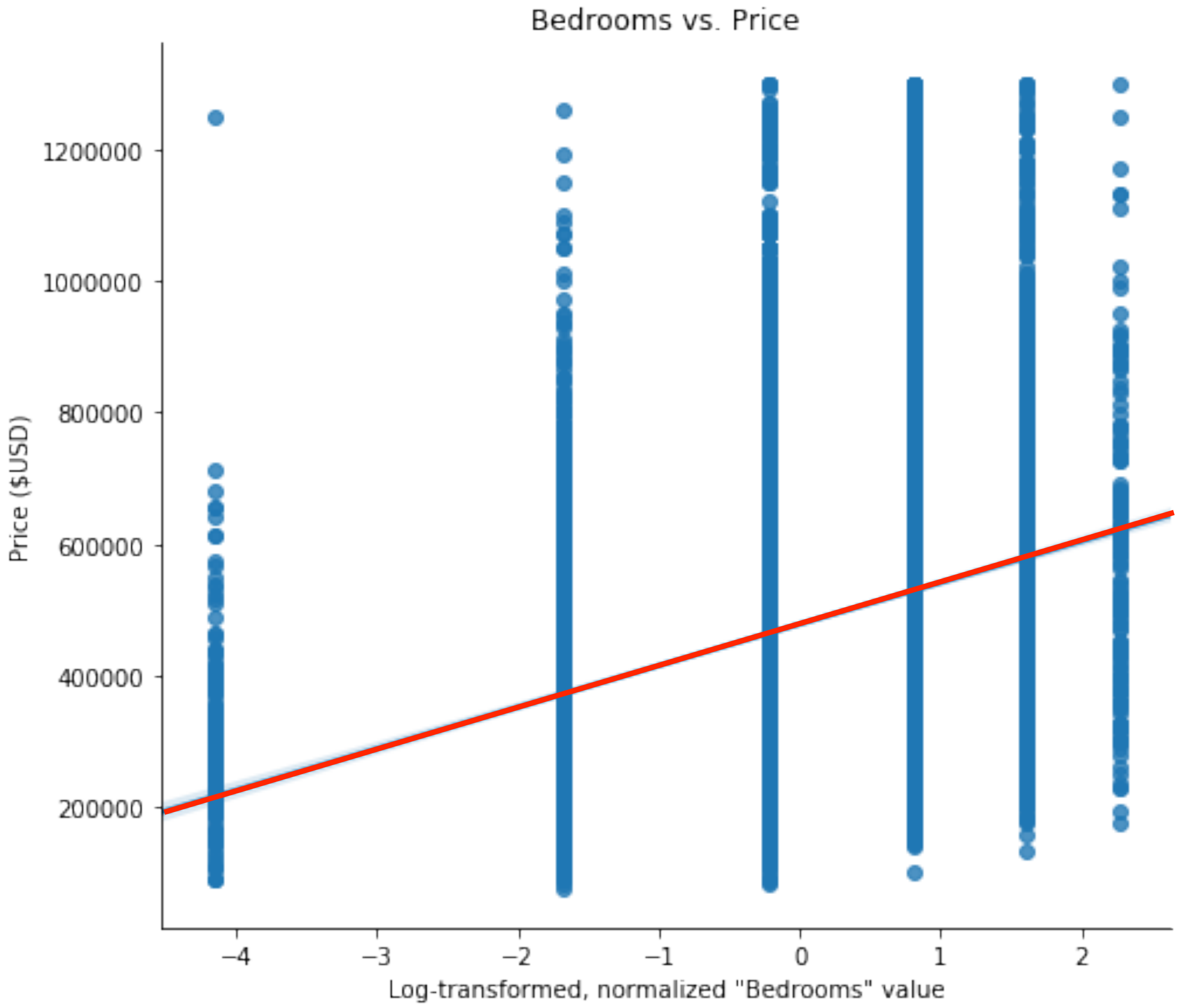
- Bump up the prices of any of our listed inventory in the most expensive zipcodes;
- Institute caps on the bids we make for investments in cheaper zip codes

3. Home square-footage and no. of bedrooms have the most impact on predicted home price.

Living square-feet:



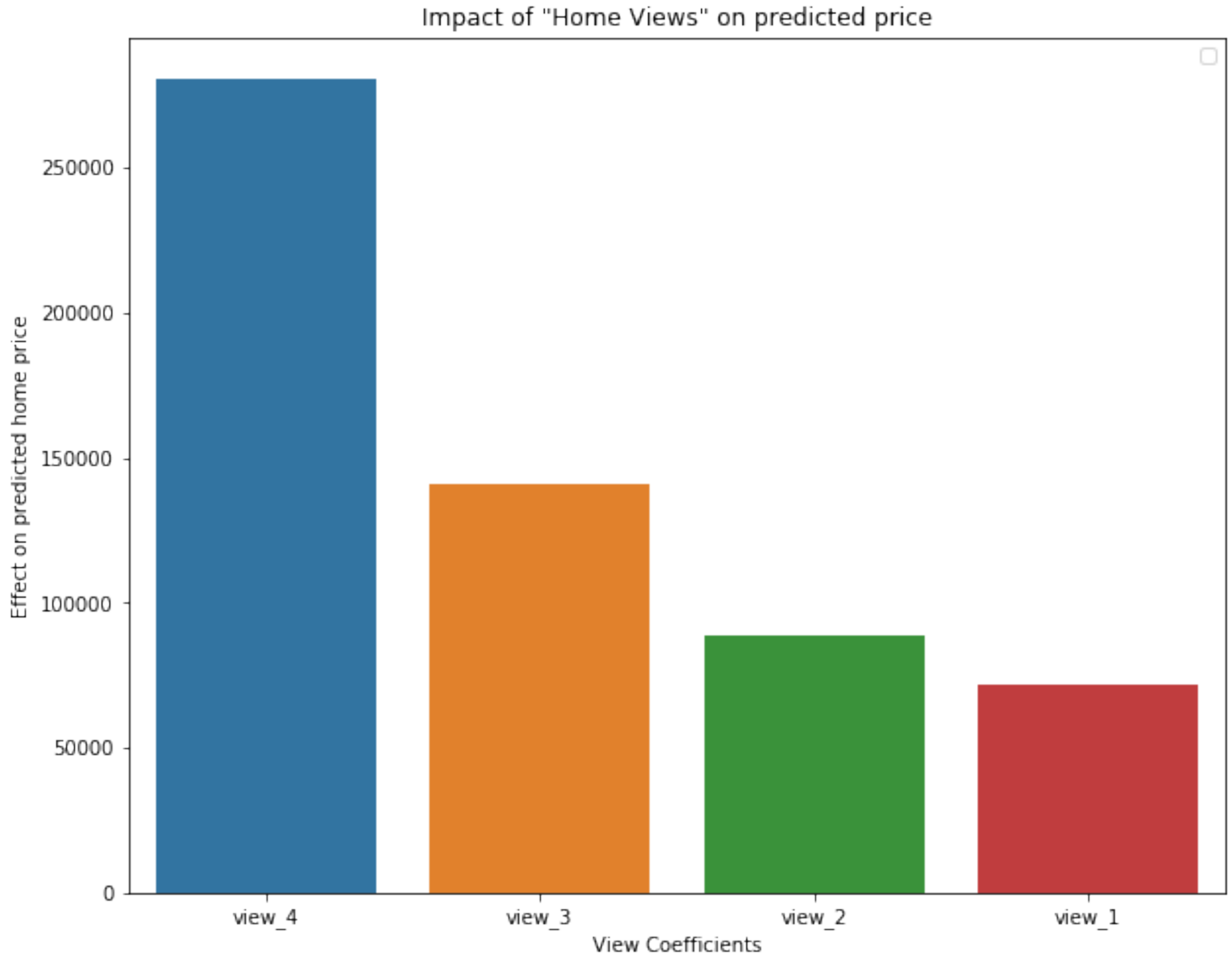
Number of bedrooms:



Recommendation:

- Home additions are the most direct way for us to improve the value of a given home in our inventory.

4. Are there any other factors that have a high impact on the predicted price of a home?

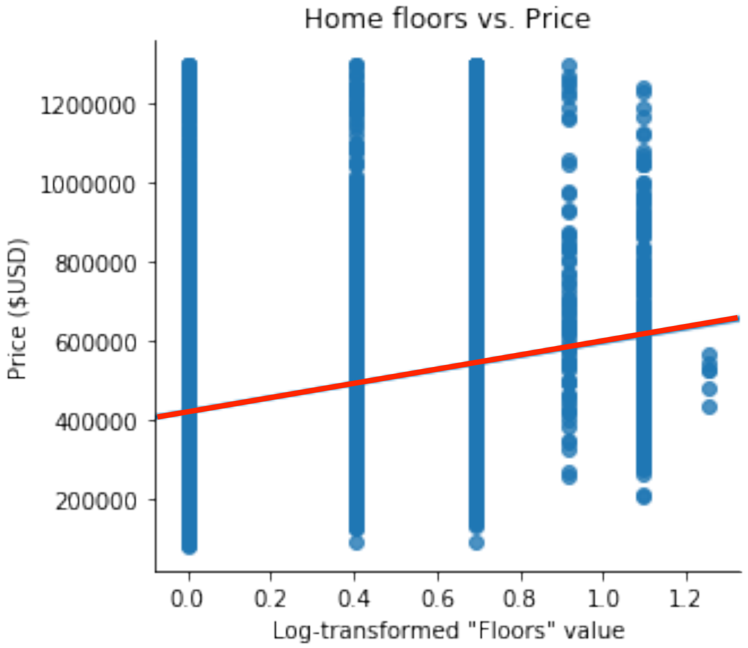


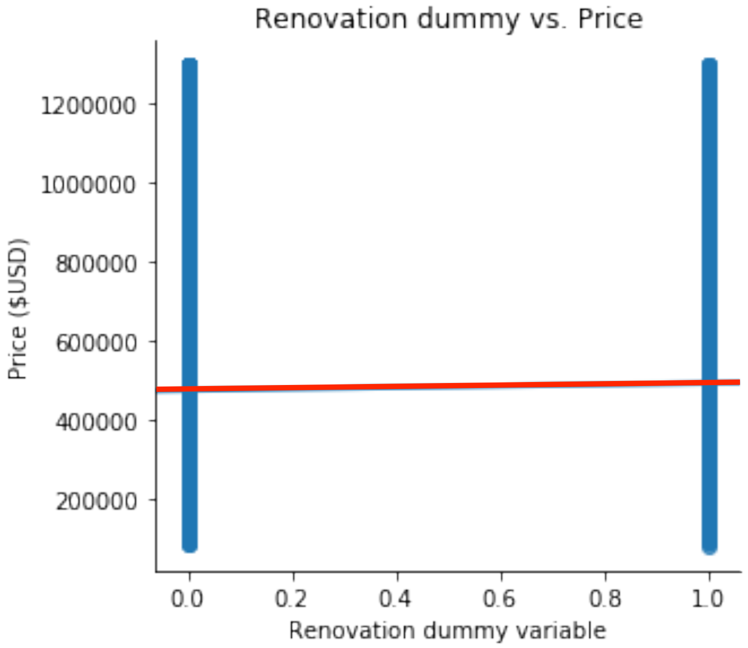
- The model adds approx. $150k to the predicted price of homes that have been viewed 3 times
- The model adds approx. $275k to the predicted price of homes that have been viewed 4 times

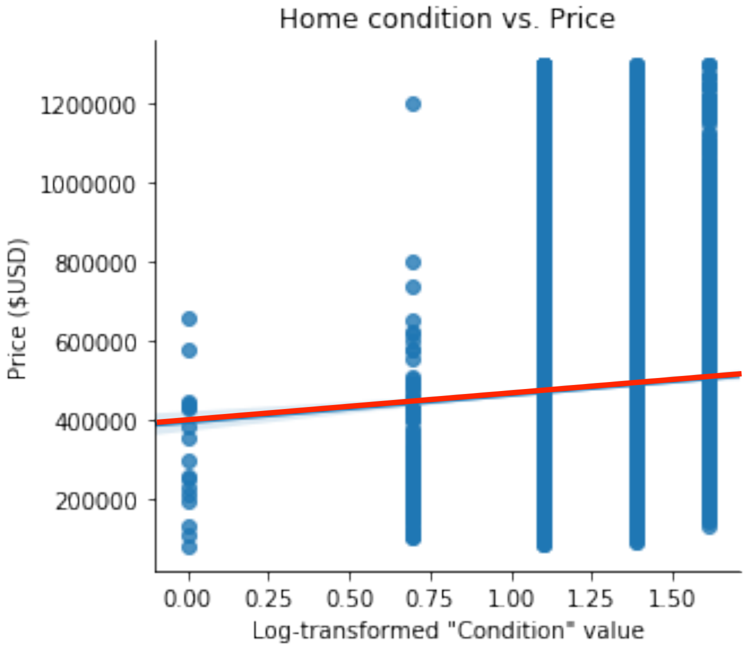
Recommendation:

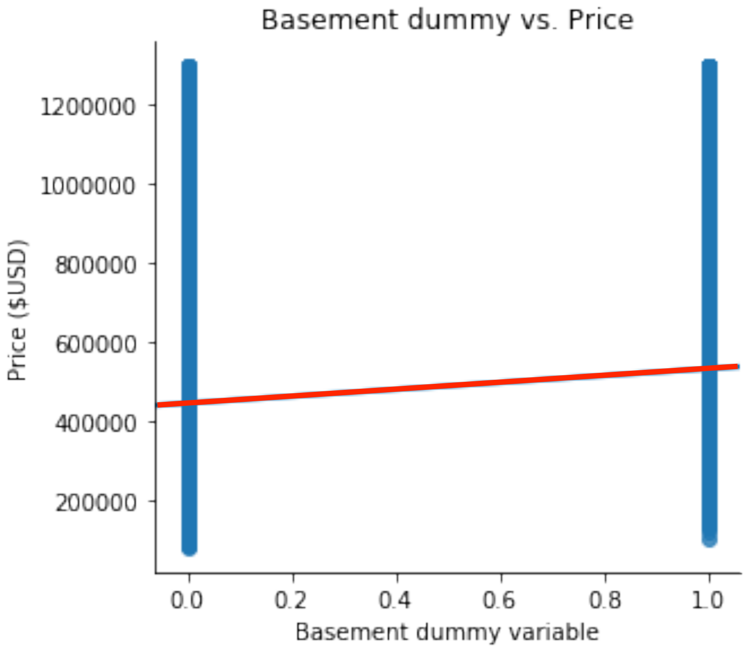
- Start showing all the homes in our inventory to immediately boost their predicted price
- Homes in our inventory should be shown 2-4 times BEFORE a final sale price is agreed upon

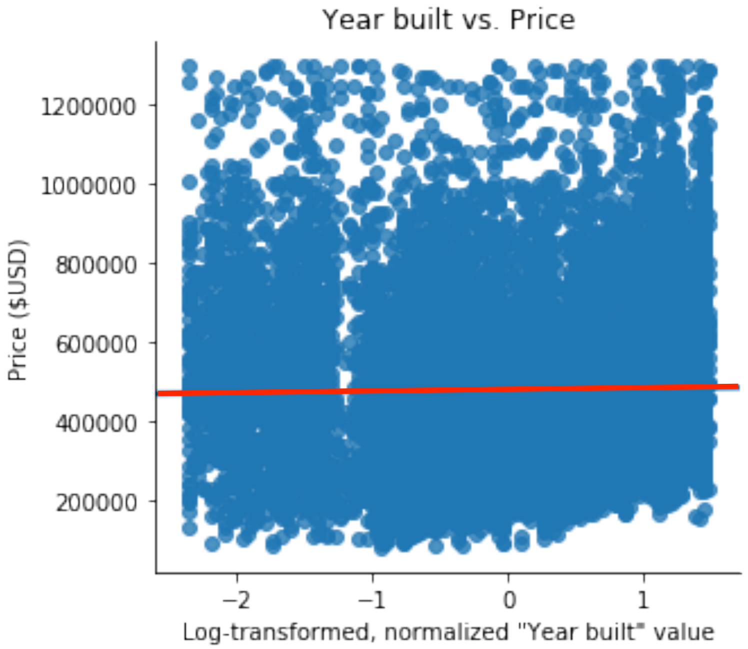
5. Appendix: Predictive power of remaining variables in our model

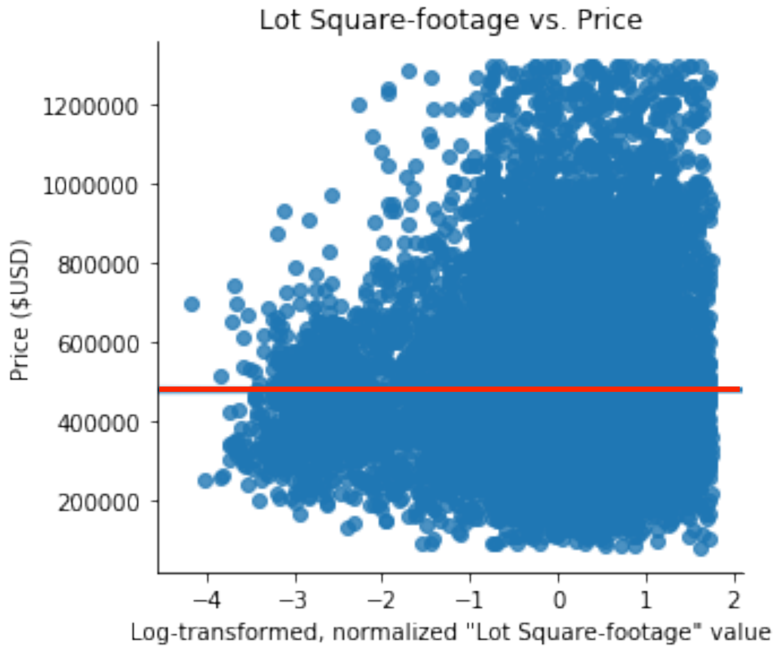

Home floors vs. Price


Renovation dummy vs. Price


Home condition vs. Price


Basement dummy vs. Price


Year built vs. Price


Lot Square-footage vs. Price

Recap of Recommendations

1. The model is useful for predicting home price with approximately $100k of the actual value of a home, but MUST be used in conjunction with our collective professional expertise and experiences for the precise valuation of our inventory.

2. Zip code is, in fact, a strong predictor of home price:
   - Bump up the prices of any of our listed inventory in the most expensive zipcodes;
   - Institute caps on the bids we make for investments in cheaper zip codes.

3. Living square-feet and number of bedrooms are the strongest value drivers of a home according to our model
   - Home additions are the most direct way for us to improve the value of a given home in our inventory.

4. The number of times a home has been viewed is also a strong driver of predicted home price
   - Start showing all the homes in our inventory to immediately boost their predicted price
   - Homes in our inventory should be shown 2-4 times BEFORE a final sale price is agreed upon

Future work

- Bring in additional home price data to further refine and optimize the prediction model

- Deeper examination of each zipcode variable individually, to ensure sufficient sample sizes have been collected and to confirm that the underlying assumptions of linear regression are upheld across the board

Thank you!

Kevin Giroux
kevinsgiroux@gmail.com