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

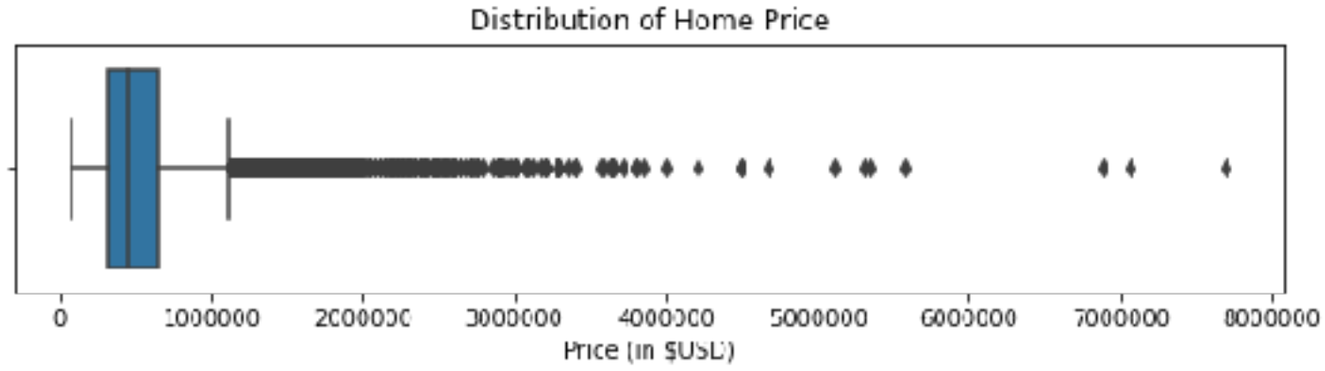
- **Secondary goal**: Provide insight into how various factors affect the predicted sale price of home, with a particular focus on the 'zipcode' variable

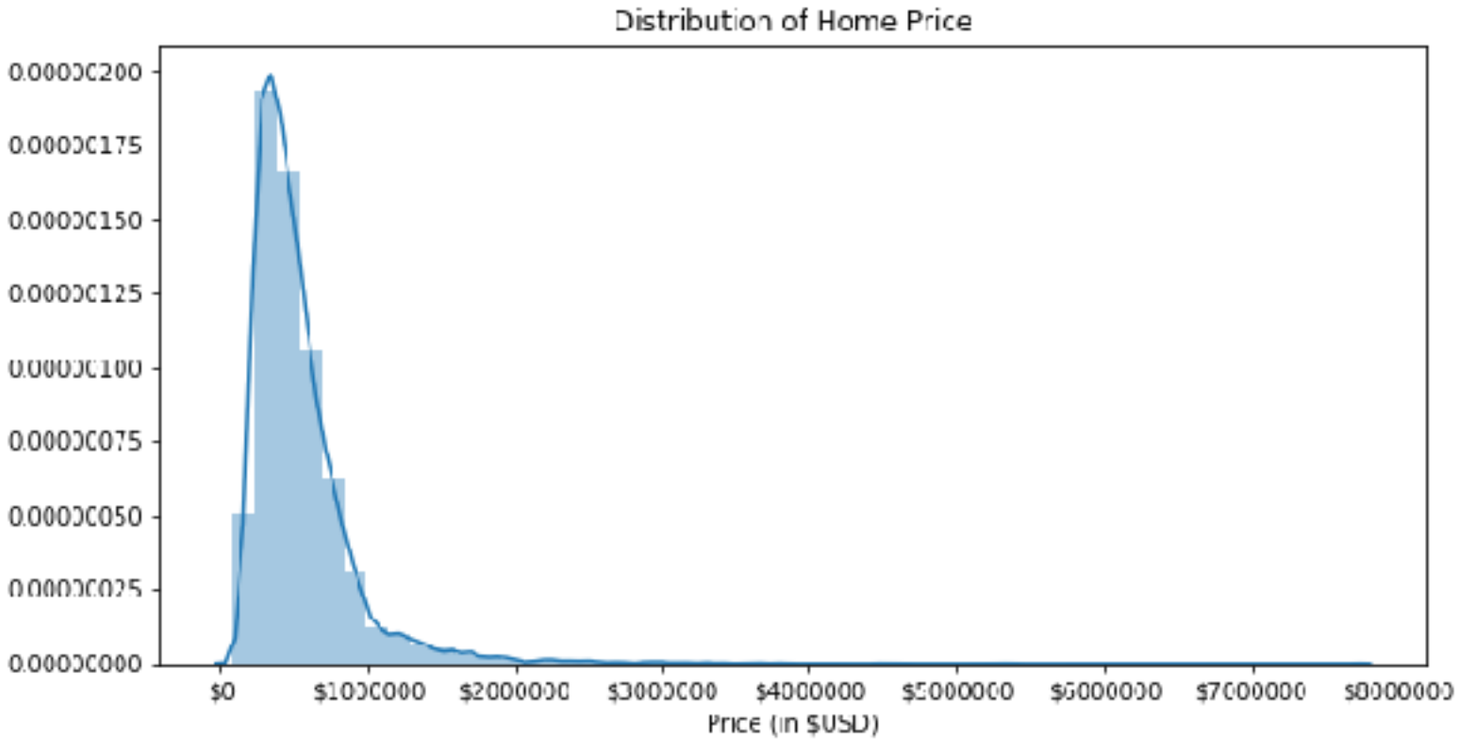
Dataset overview

- 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 rooms in home with a view)
  - 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

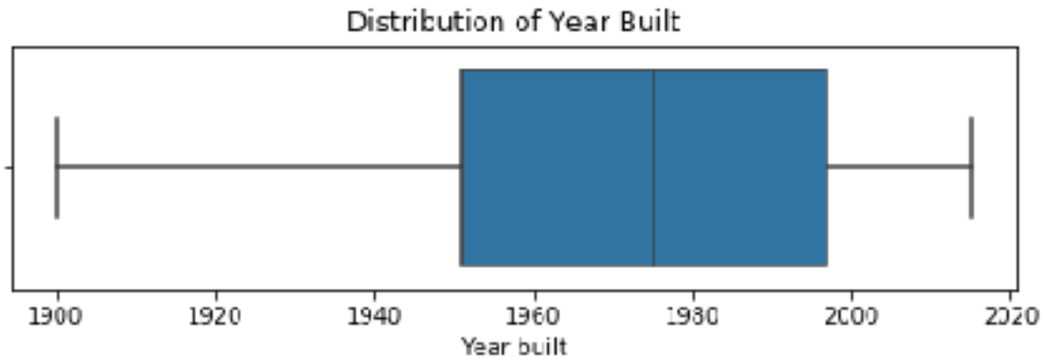
Min: $78,000

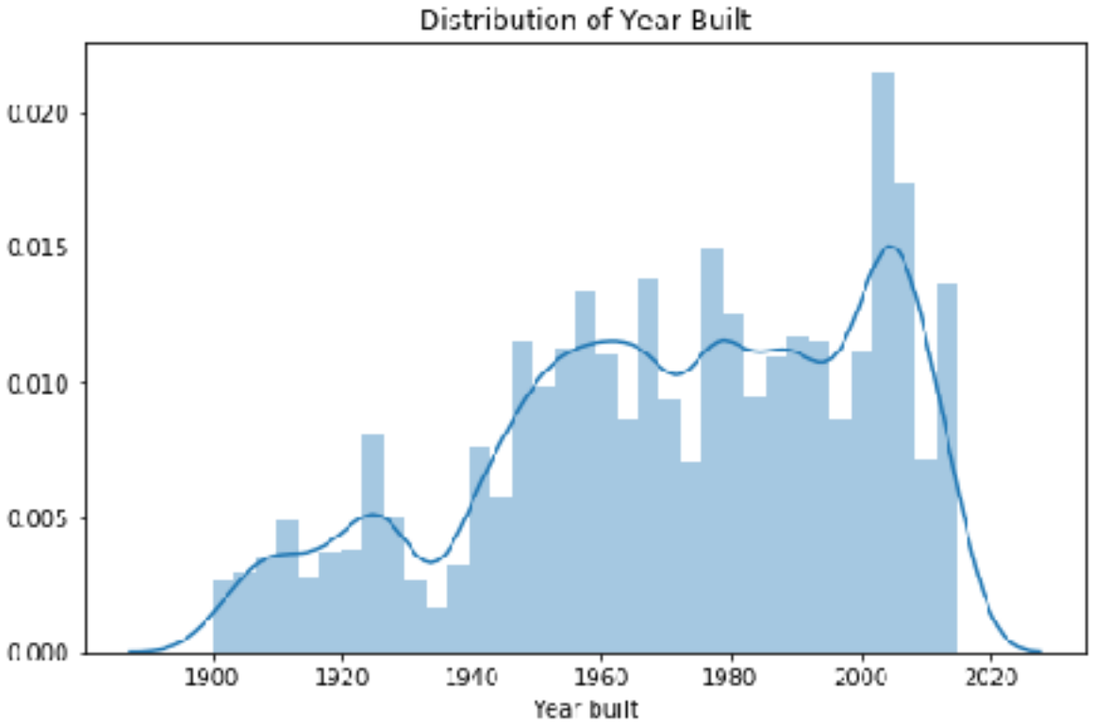
Max: $7,700,000

Conclusion: Outlier removal necessary prior to building a predictive model

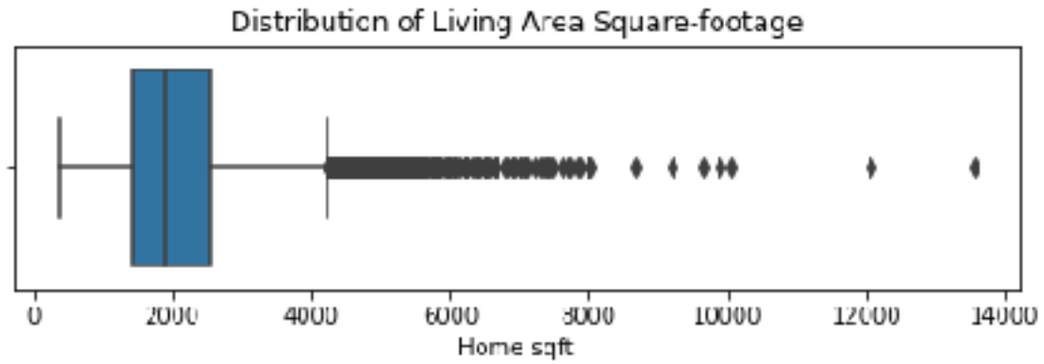
Dataset overview

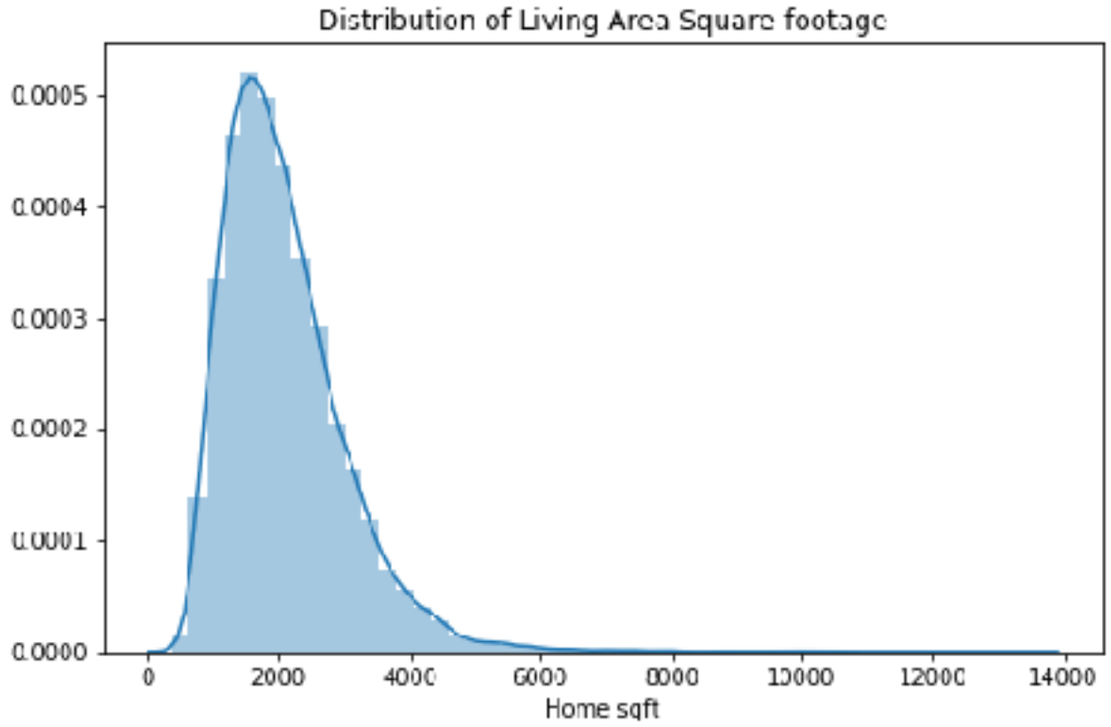
YEAR BUILT:





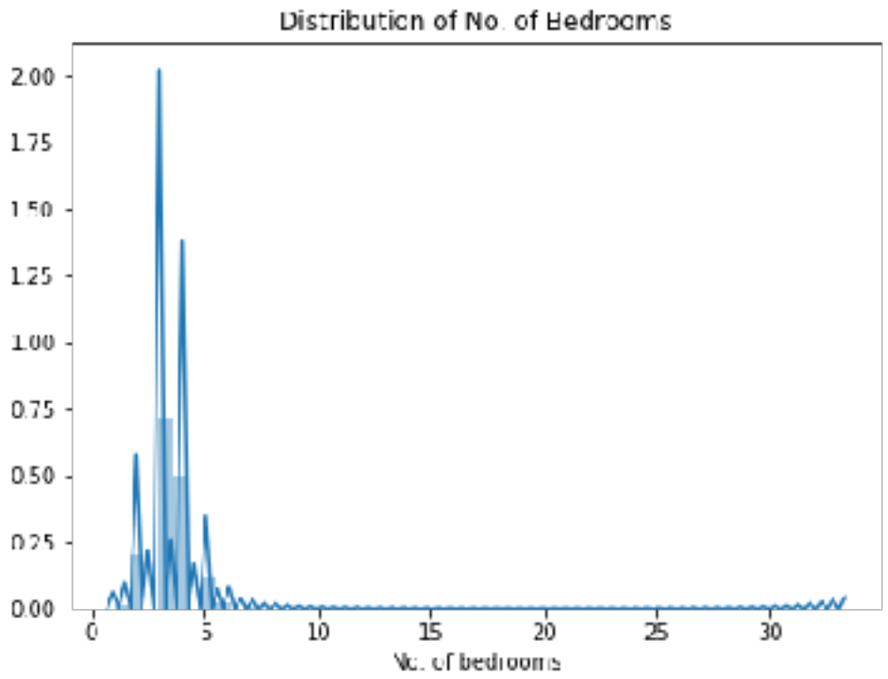
HOME SQFT:



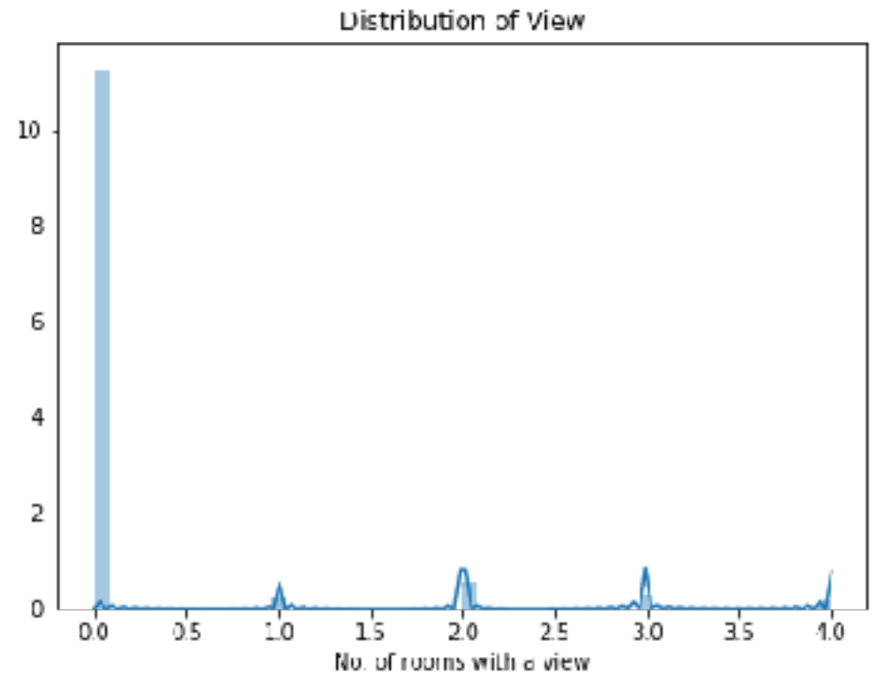


Dataset overview

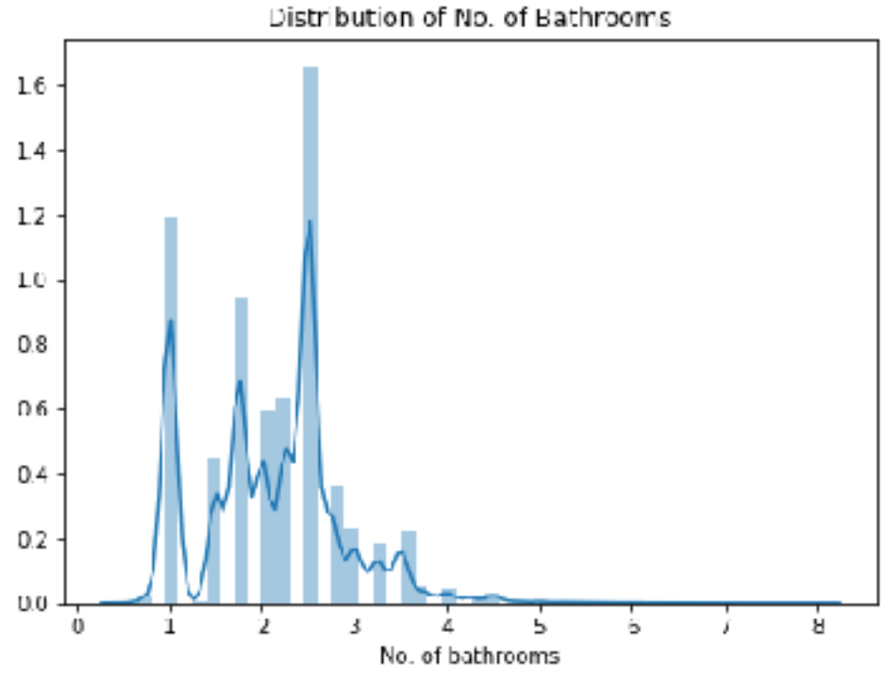
BEDROOMS:



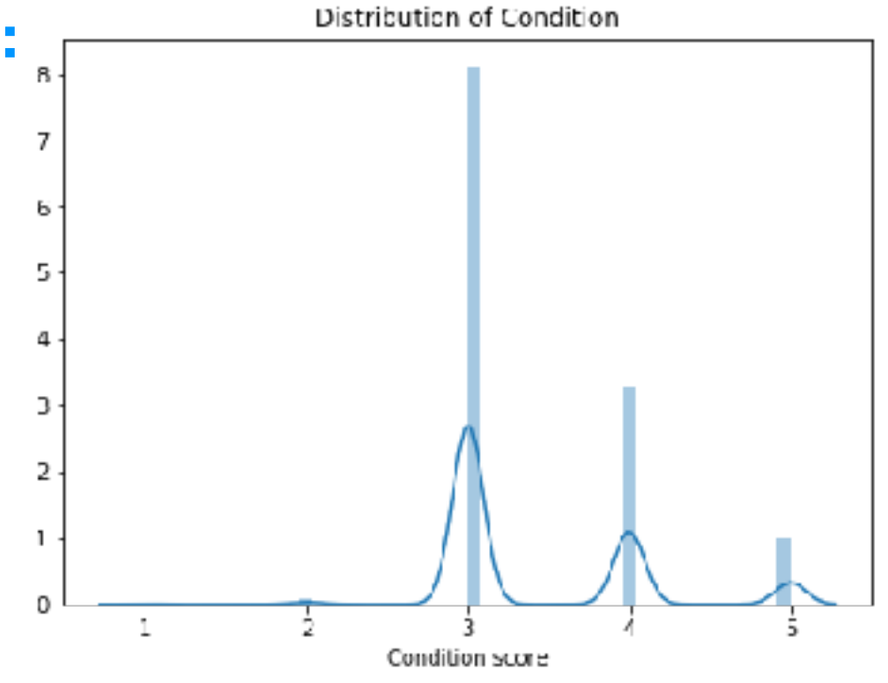
NO. OF ROOMS W/ A VIEW:



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

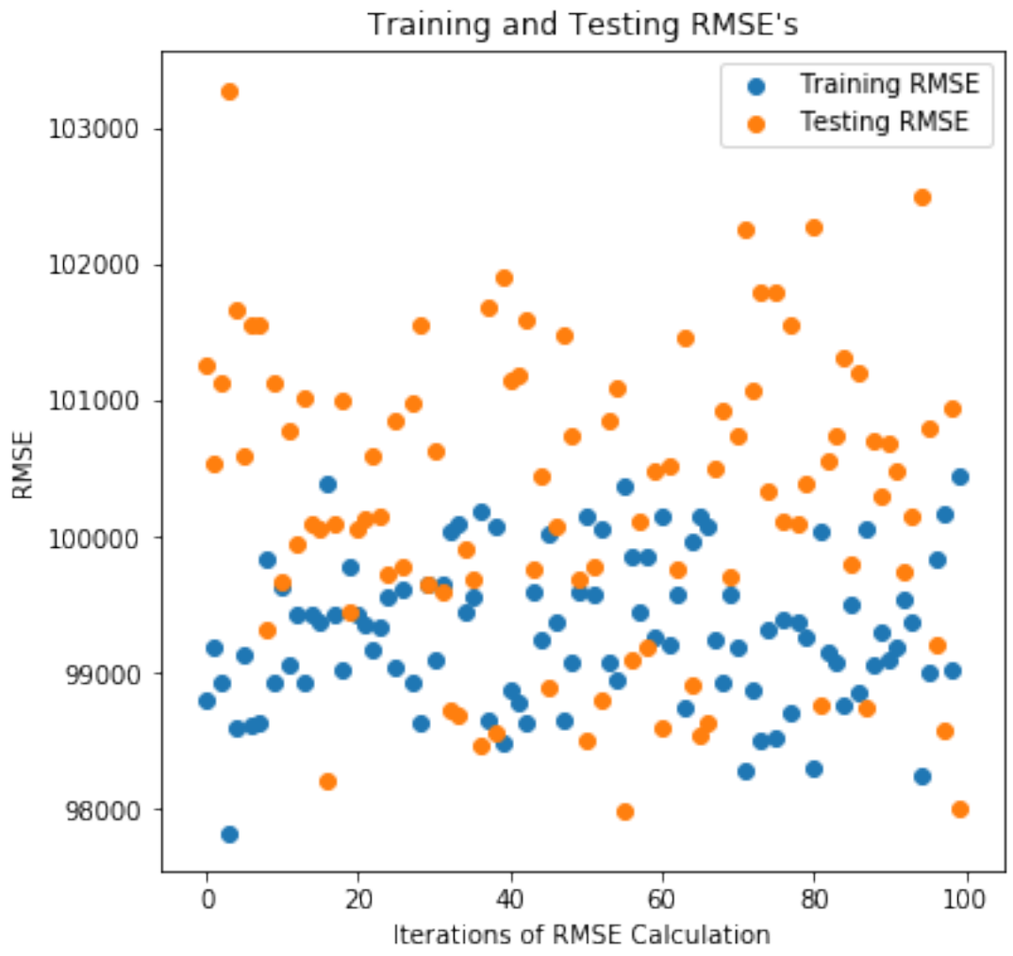
- Analysis of each feature in the sample data, and removal of outliers
- As a result, our model was trained to best predict prices of homes with the following characteristics:
    - Home price <= $1.3mm
    - Bedrooms <= 6
    - Living sqft <= 4500
    - Lot sqft <= 17,500
- Trained a multi-linear regression 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?
- How does home square-footage affect the predicted price of a home?
- Are there any other factors that have a high impact on the predicted price of a home?

RESULTS

1. How accurately can the model predict home price?



- The standard deviation of the model's predicted price around a home's actual price is approximately $100k

- In other words, 68% of the time, the model predicts home price within $100k of the real-world price

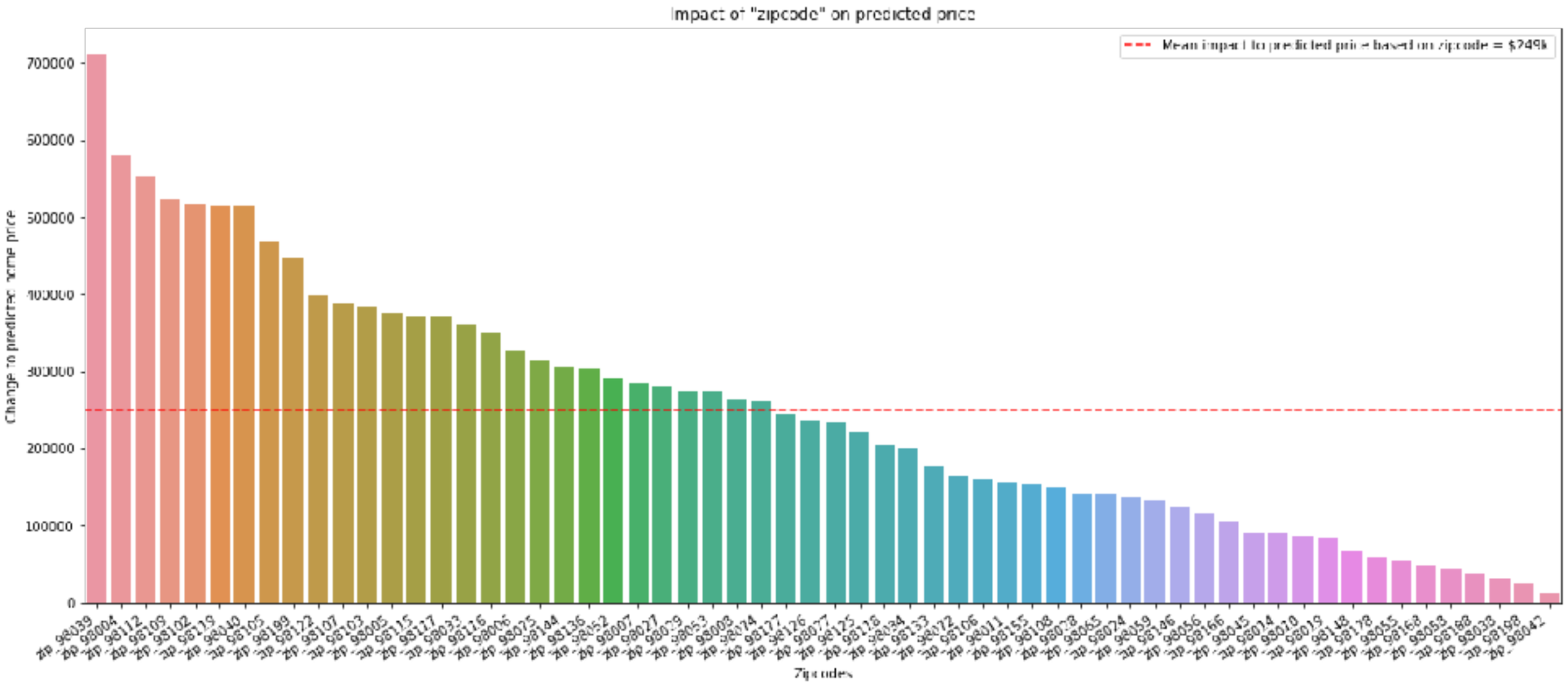
NOTE: RMSE stands for "root mean square error" and is a measure of the difference between the model's predicted home price and the actual home price

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

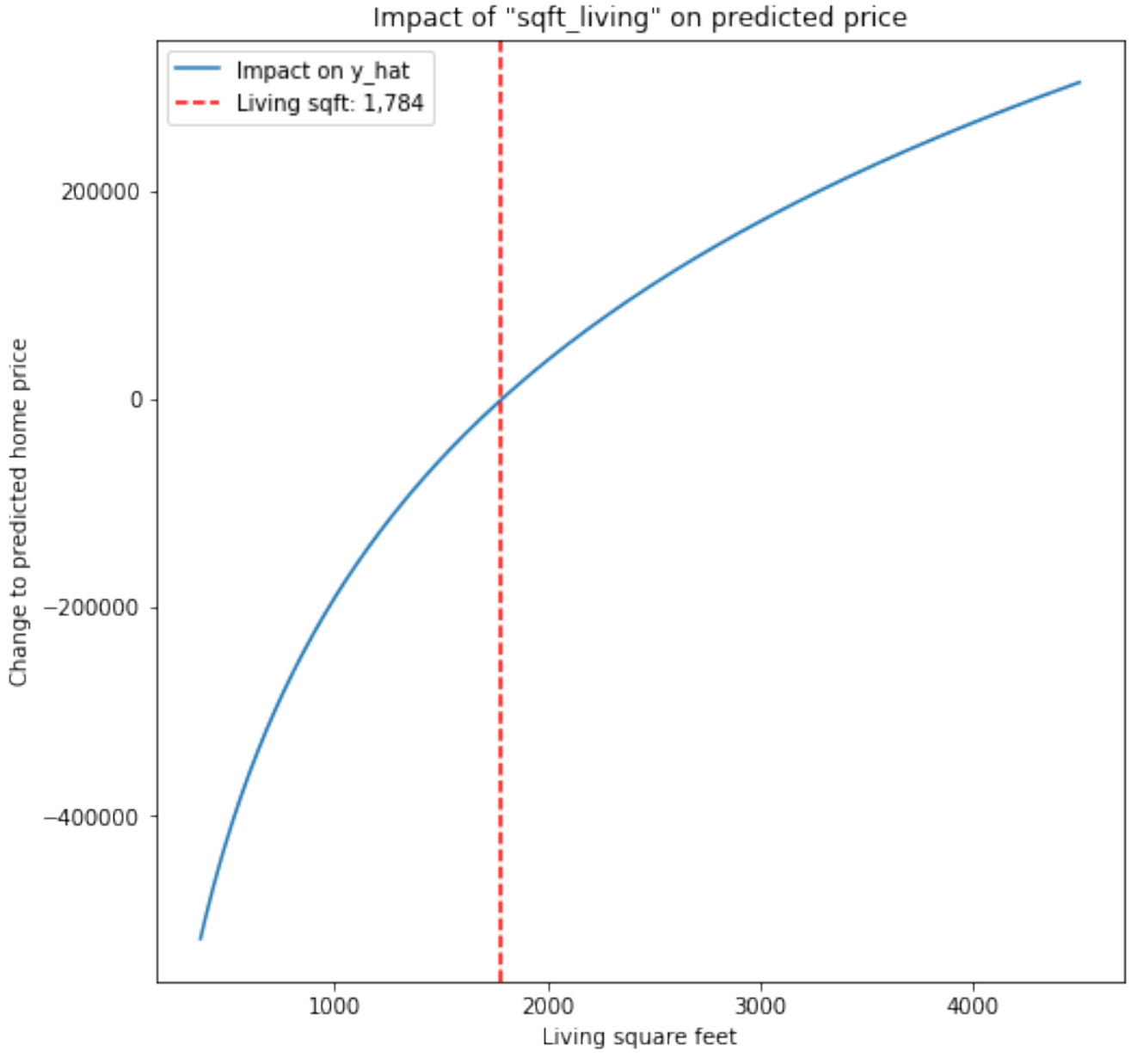
Recommendation: Group and price homes in our inventory by zip code

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



Recommendation: Group and price homes in our inventory by zip code

3. How does home square-footage affect the predicted price of a home?



- The model assumes a hypothetical baseline living space square-footage of **1784 square feet**

- **Homes = 1784 square feet:** the model would not add or subtract any amount to the price of the home (as predicted based on the other independent variables included in the model)

- **Homes > 1784 square feet:** the model incrementally **adds to the home price** as predicted by the other variables.

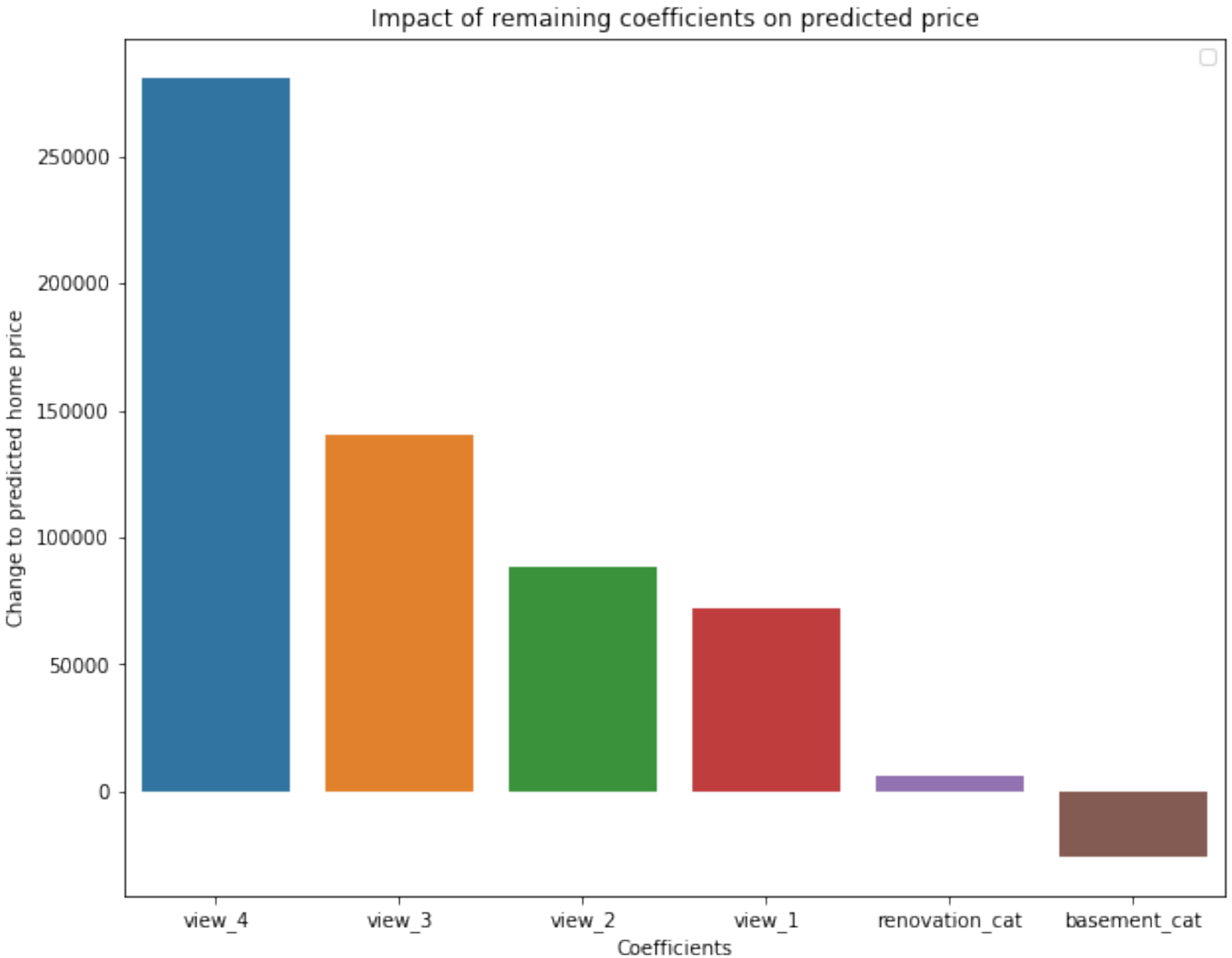
- **Homes < 1784 square feet:** the model incrementally **subtracts from the home price** as predicted by the other variables.

3. How does home square-footage affect the predicted price of a home?

Recommended next step: Identify the square-footage at which point we begin to see decreasing marginal increases in the predicted price.

In other words, identify what square-footage home is the best deal (according to our model).

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 with 3 rooms with views

- The model adds approx. $275k to the predicted price of homes with 4 views (double!)

- The model adds almost nothing to the predicted price of a home for having been renovated

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

Recommendations:

1. Identify investment homes where a fourth view could be easily created, as this dramatically improves the predicted home value

2. Don’t over-invest in renovations, which our model does not place a ton of weight on when predicting price

Recap of Conclusions

1. The model can be used to predict home sale price within $100k of the actual sale price of a home

2. Zipcode is, in fact, a strong predictor of home price; we recommend grouping and pricing homes in our inventory by zip code

3. Our model adds to the predicted price of homes greater than 1,784 square feet; we recommend identifying the “best-deal” square footage where the value added by the model for each additional square foot of living space begins decreasing

4. Number of rooms with views is also a strong value driver; look for candidate investment homes where a fourth window with a good view could be easily installed

5. Other factors significant to the prediction of home price, which the firm should keep in mind, include lot square footage, the presence of a basement, home condition, number of bedrooms, number of floors, and year built

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