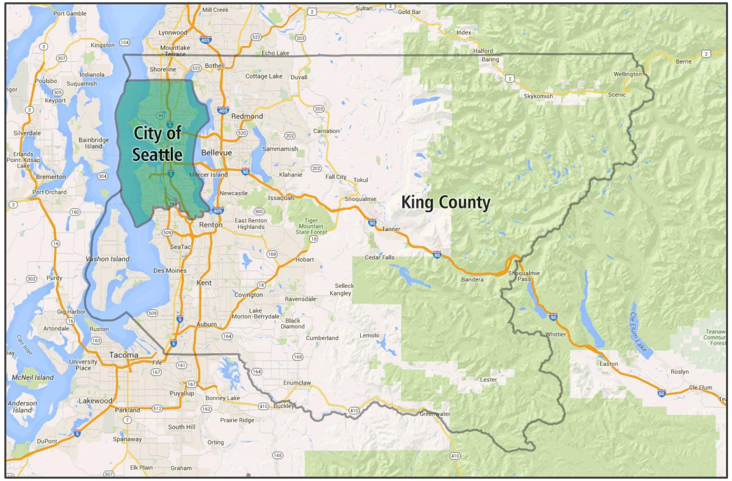
Module 2 Final Project

Will Dougherty
Flatiron School
Data Science, self-paced

King County, WA

Predicting House Prices



Dataset used:

House data from May 2014 - May 2015

Business problem:

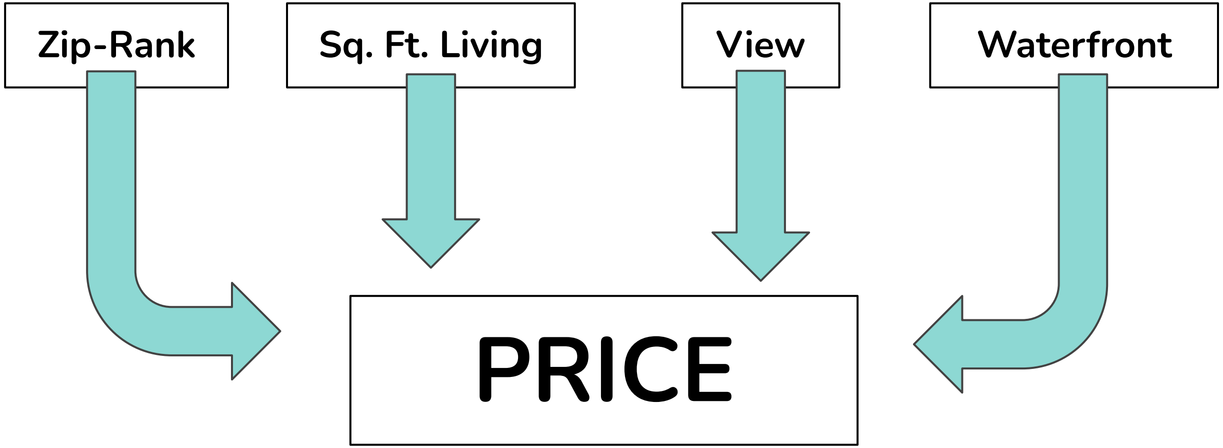
For our new startup **website**, a local house-listing website for those looking to buy, as well as sell houses:

How can this data be used to model house prices to create a **prediction engine** for **users** looking to **list** and **sell** their house?

We want to find the **best features**, keep it **comprehensible**, and generate **actionable insights** to inform our new tool.

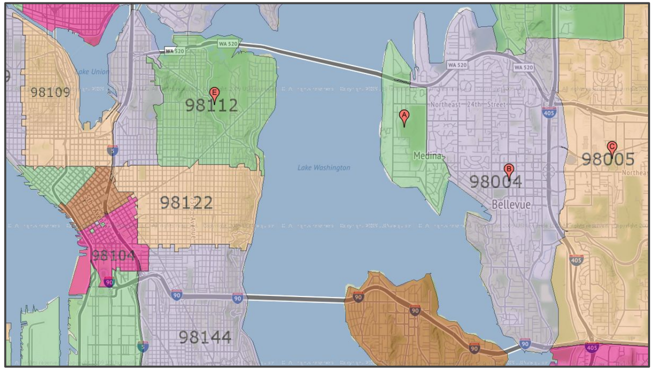
My Model

Using multiple linear regression, I've used a few different features to predict house prices, and achieved an R-squared value of ~77% - that is, **77%** of the **variation** in the data is accounted for by this model. This model is most accurate for house values **between $150,000 and $1.5 million.**



Feature Details

What is **Zip-Rank?**



Using the **median home price** for each **zipcode**,
I assigned a value using this formula:

$$\frac{\text{Median Home Price of a given zipcode}}{\text{Highest Median Home Price (\$1,895,000)}} = \text{Zip-Rank}$$

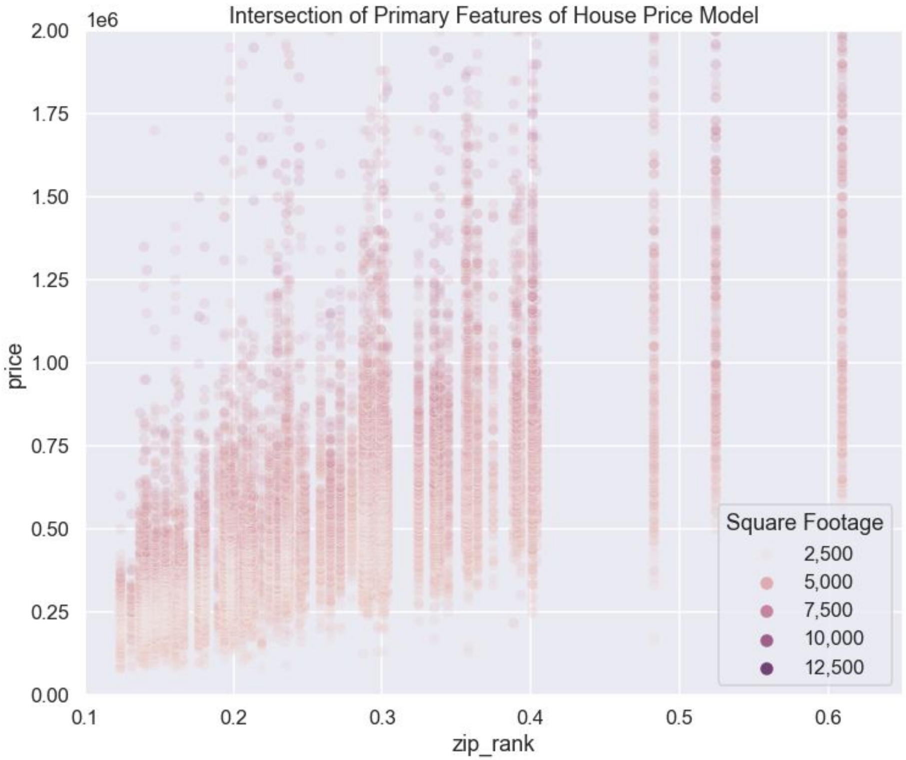
This provides a baseline value for any home in that zipcode,
to be modified by square footage and view/waterfront values.

Thus, the highest rank is 1.0 (for that highest-median zipcode) and the lowest is 0.12

The median zip-rank is 0.235, which represents a median home price of ~ $446,000

Zip-rank vs. Sqft. Living

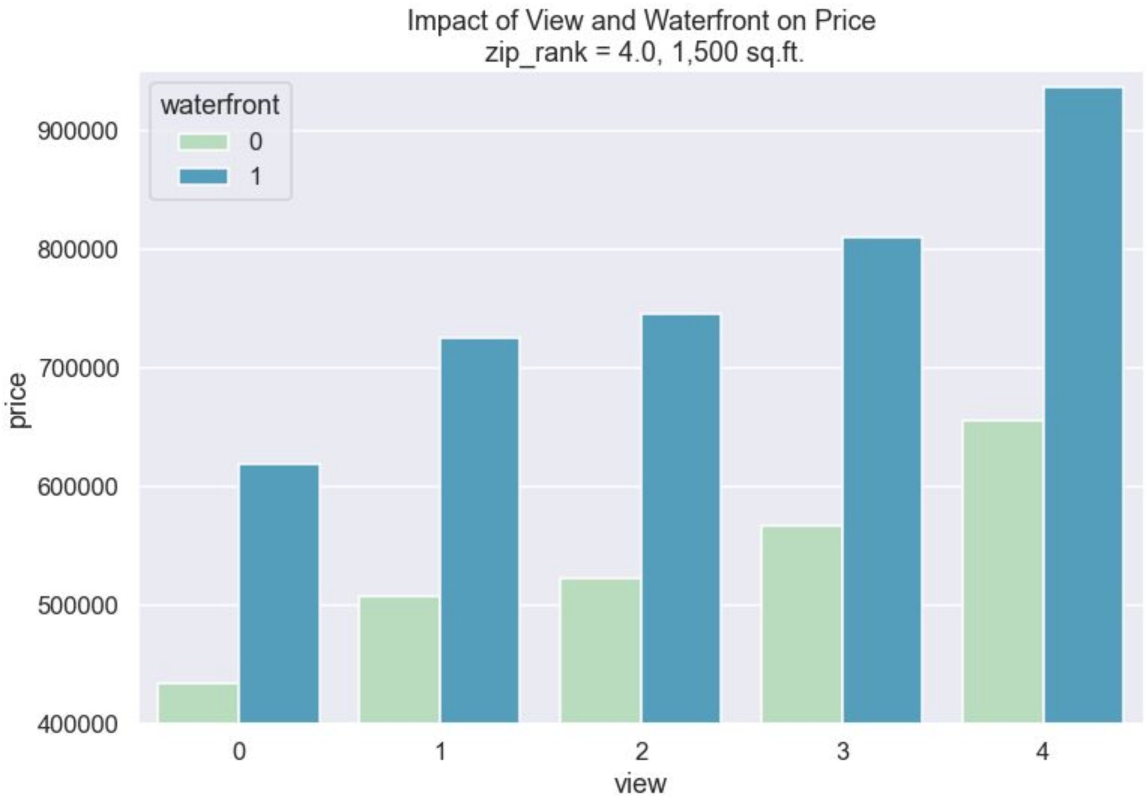
Sqft. Living -
square-footage of ‘living’ space (non-basement), divided by 1,000 (1,200 sq.ft. = 1.2)



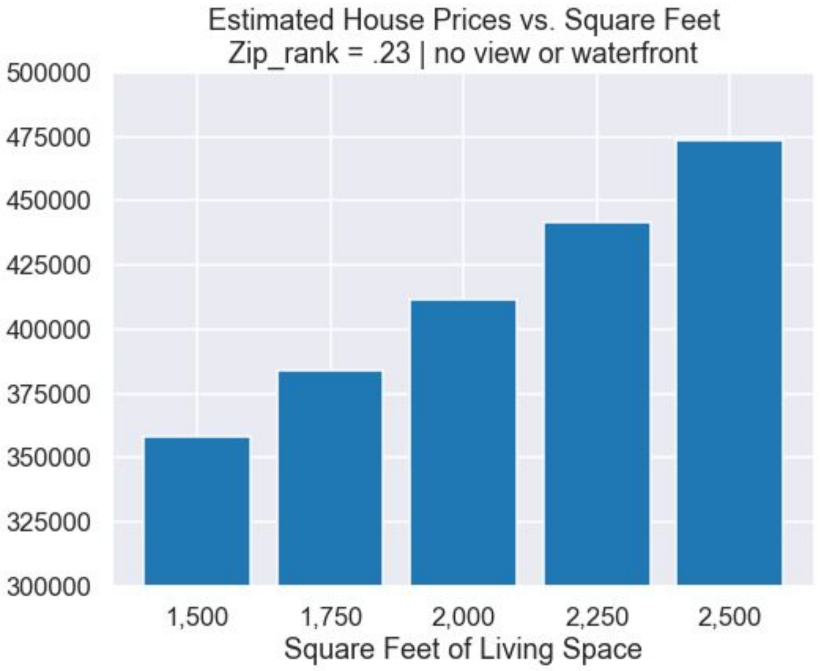
Impact of View and Waterfront

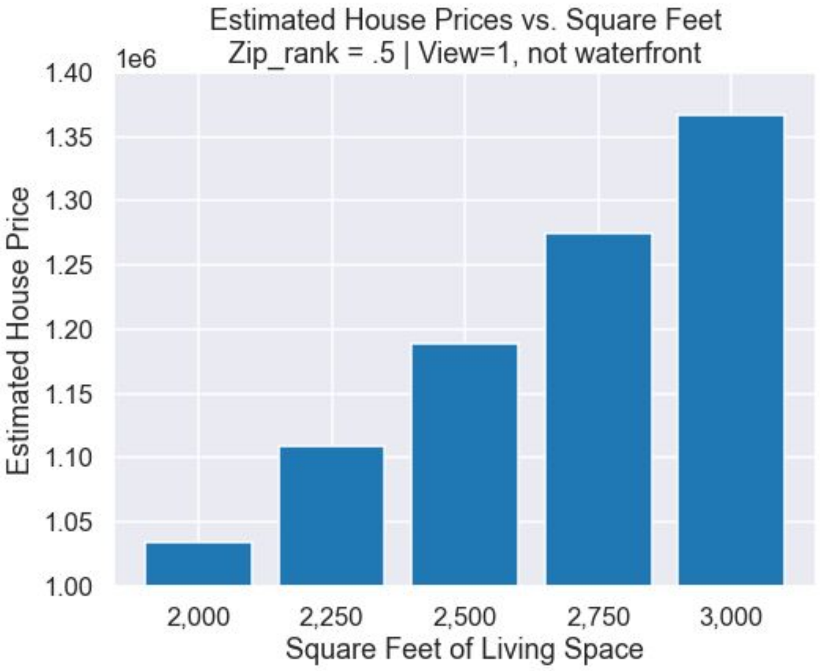
View:
Ranges from 0 - 4
(no view to amazing view)

Waterfront:
Is either 1 or 0
(is on the water or not)

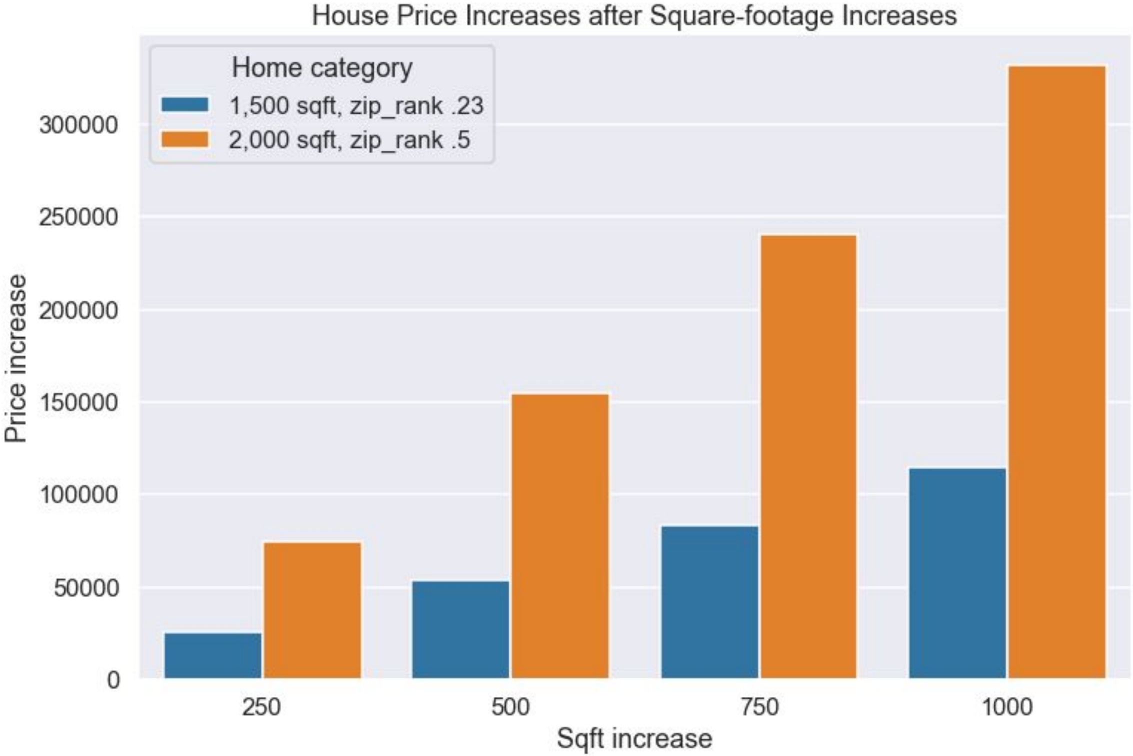


Example square-footage / house price predictions:





As baseline values go up, value gained by increasing square footage increases:



Recommendations

1. Using this model, a tool could be developed in addition to the house price predictor: a **recommendation engine**, to show how much a user's house would be worth if the square footage increased by certain amounts. This could be supplemented by a **calculator** for determining if certain **renovations/additions** would be worth investing in before selling the house.

2. **Zip-rank** can be kept relevant and accurate by continually updating median house prices for each zipcode, and adjusting the model accordingly.

3. Since **view** and **waterfront** values are somewhat abstract, determining the parameters for classifying those values could help users determine their own view/waterfront values for their home. This could be a tool that has example images, or it could suggest values based on known values in their neighborhood (since if a house has a great view, other houses on the same street would likely have the same or similar view).

4. Model **caveats**: for very low and very high predicted house prices, accuracy is less. If a user's house is predicted to be lower than ~$150,000 or greater than ~$1.5 million, the user needs to be informed of this and given other options to determine their house's value.

Future Work

- As mentioned, keeping **zip-rank** relevant over time is essential.
- Due to limitations of this modelling technique, things like **grade**, **bedrooms/bathrooms**, **basement** information, and **lot size** were largely either not significantly contributing to the model, or were redundant with **sqft-living**. Finding other methods of incorporating these features could help further increase prediction accuracy.

Thank you!