Predicting Home Listing Prices in Vermont

Andrew Smith

Goals

- Understand features that contribute to home listing prices in Vermont
  - A tool for real estate investors to identify undervalued properties
  - A tool for agents to guide their clients
  - Guide home remodelling (i.e. how much will the new bathroom increase the home price

Tools

Web Scraping

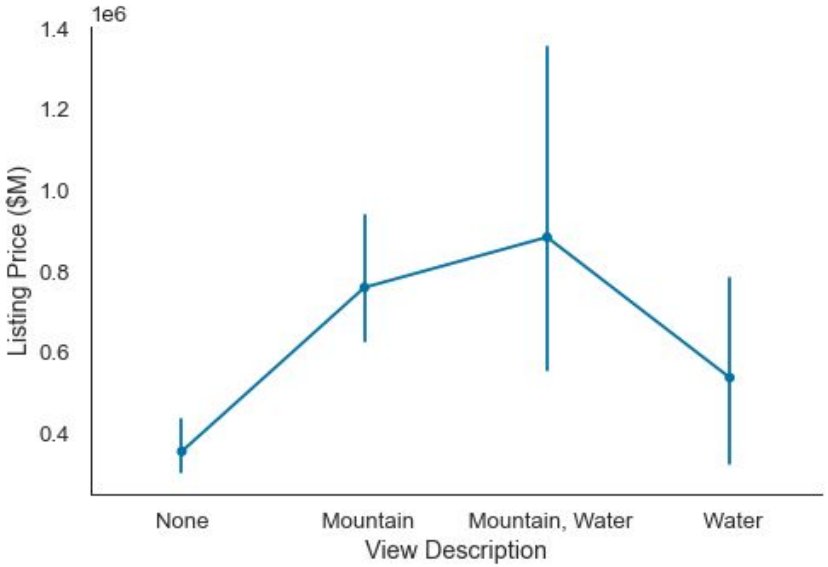
Modelling / Cleaning / Viz

Data Sources & Data Description

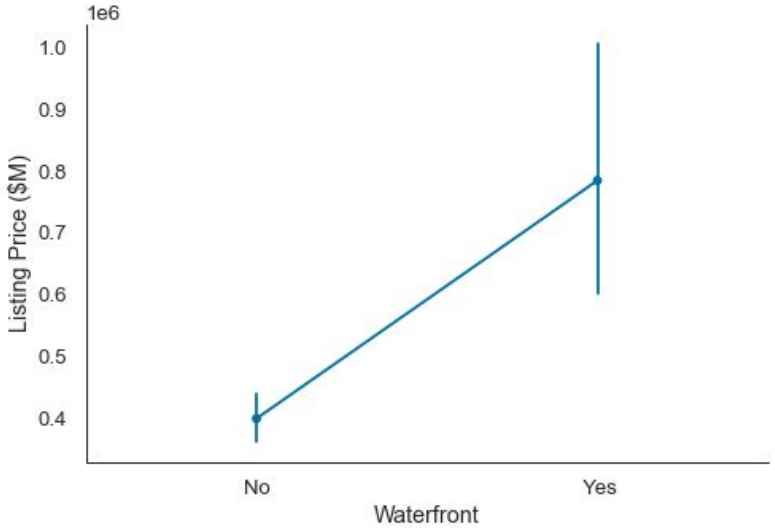
- Data sources:
  - All VT single-family homes for sale on *Zillow* (729 total)
  - County level median income from the *U.S Dept. of Housing and Urban Development*
  - Created County to Zip Code Crosswalk
- Variables Scraped from Zillow:
  - Dependent variable:
    - Listing Price
  - Features *(11)*:
    - Number of Bedrooms
    - Number of Bathrooms
    - Year Built
    - Lot Size (sq ft.)
    - View Description
    - On Waterfront
    - House Style
    - New Construction
    - Garage Spaces
    - Zip Code
    - County Median Family Income
    - Zip Code

Exploratory Data Analysis: Categorical Variables

Listing Price ($M) vs. View Type



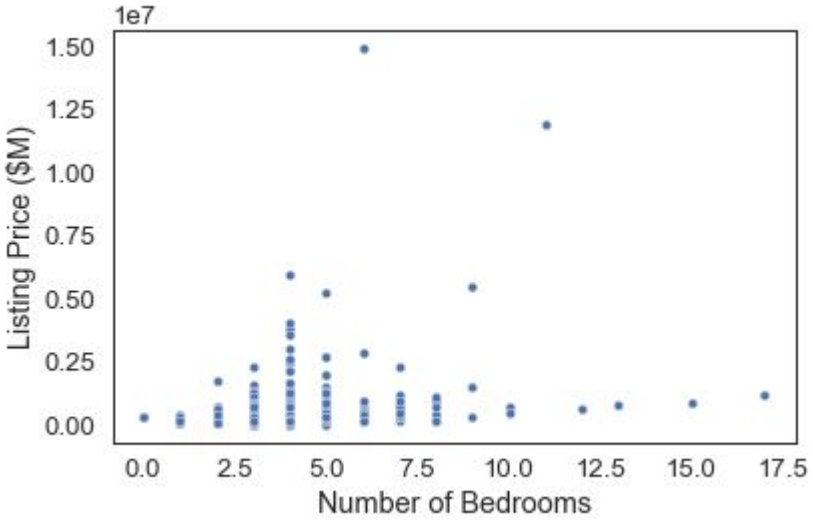
Listing Price ($M) vs. Waterfront



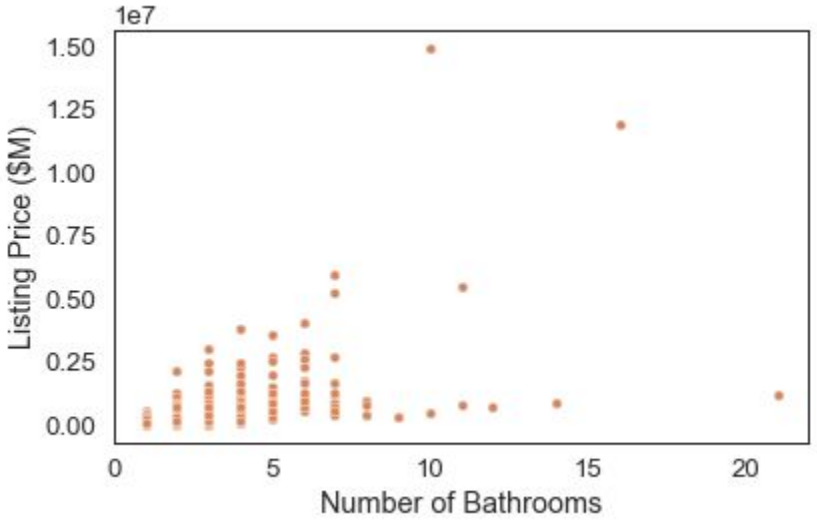
- Other categorical variables (year built, new construction, house style) had limited visual correlation

Exploratory Data Analysis: Continuous Variables

Listing Price ($M) vs. Number of Bedrooms



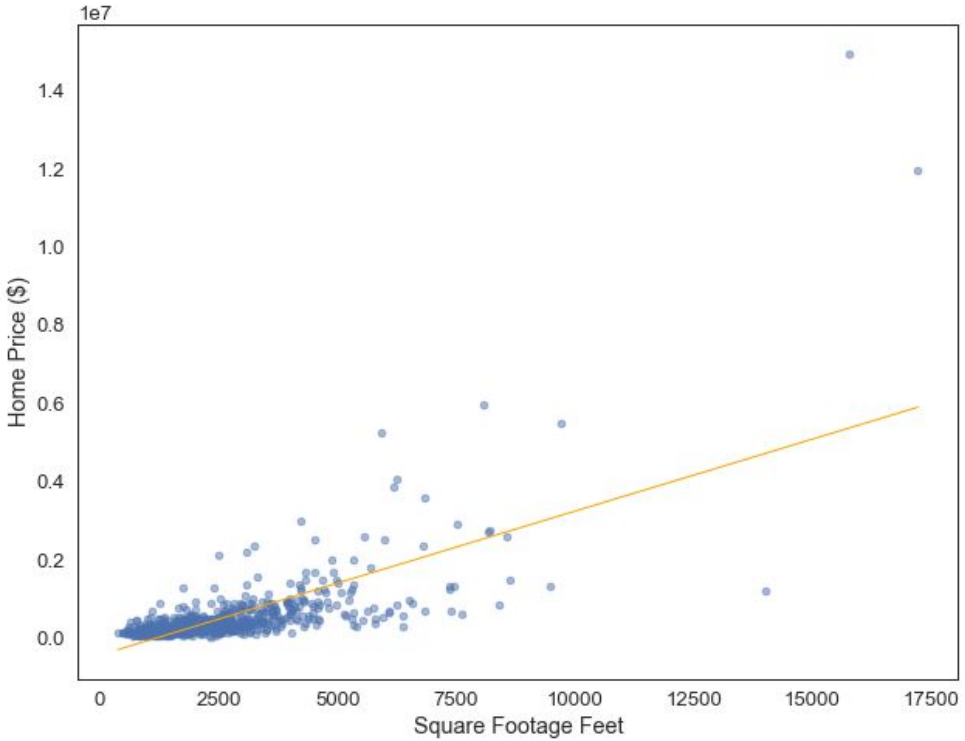
Listing Price ($M) vs. Number of Bathrooms



- Homes with a larger number of bedrooms and bathrooms and throws off a second order or linear relationship; many of these homes are older larger homes and an interaction term with these variables helps, but ultimately still renders them to have minimal impact on the model

Simple Initial Models: Square Footage

Listing Price ($M) vs. Square Feet



Results (SF)

R^2 = 0.24

SF coefficient = 384

Results (SF^3)

R^2 = 0.34

SF^3 coefficient = 0.04

Results are validation scores

Utilizing Lasso

- Due to a small dataset, a Lasso cross-validation was unable to effectively regularize the model; however, it was used to help suggest key interaction terms for the next iteration of the model

On Waterfront (Yes / No)

Bathrooms **X** Year Built

Home Square Feet **X** lot size (sq ft.) ^ 2

View Description (Yes / No)

Home Square Feet ^ 3

Bathrooms ^ 3

On Waterfront **X** Home Square Feet

On Waterfront **X** Bedrooms

View Description **X** Bathrooms

Home Square Feet **X** lot size (sq ft.) ^ 2

On Waterfront **X** View Description

Zip Code

Median Family Income

Number of Garages

House Style

New Construction

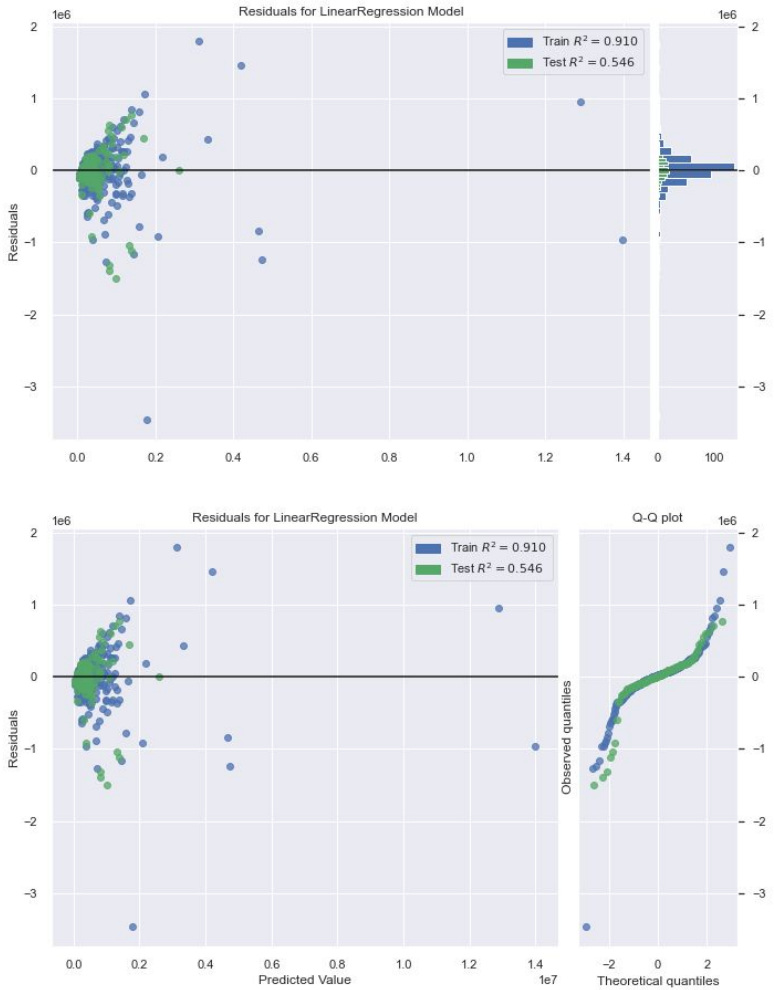
Best model results

Results

$R^2 = 0.55$

RMSE = 313,308

Coefficients: model has become too complex to interpret easily



Model Weakness / Next Steps

- Some measure of quality / last renovation
  - Residuals show that homes of lower quality with limited renovations are overvalued due to a high square footage for instance
    - Achieved through machine learning on photos to determine quality or scraping a new website
- More data points for effective regularization
  - Would try for the last few years of data / 3,000 + data points
- Better screening of single family homes
  - Various inns (marketed as SF homes) were included and had high residuals
- Better screening of “waterfront”

Final Conclusions

- SF and on waterfront hold the largest impact on house price, but are likely subject to large variation in prices depending on home quality