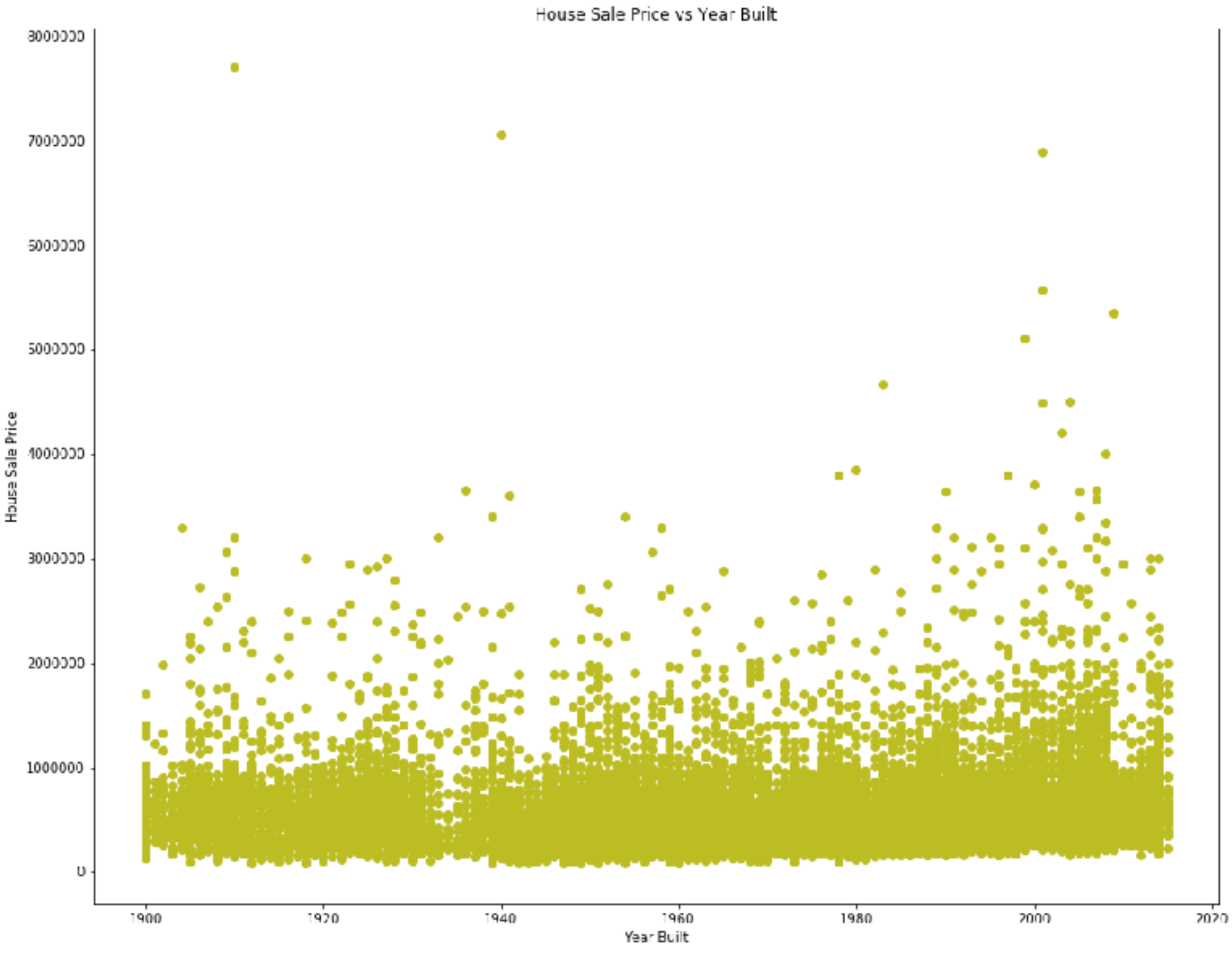
Predicting Housing Prices in King County, WA

Daniel Martinez, Jenna Cicardo, and Zach Timmons
(aka Haus Masterz)

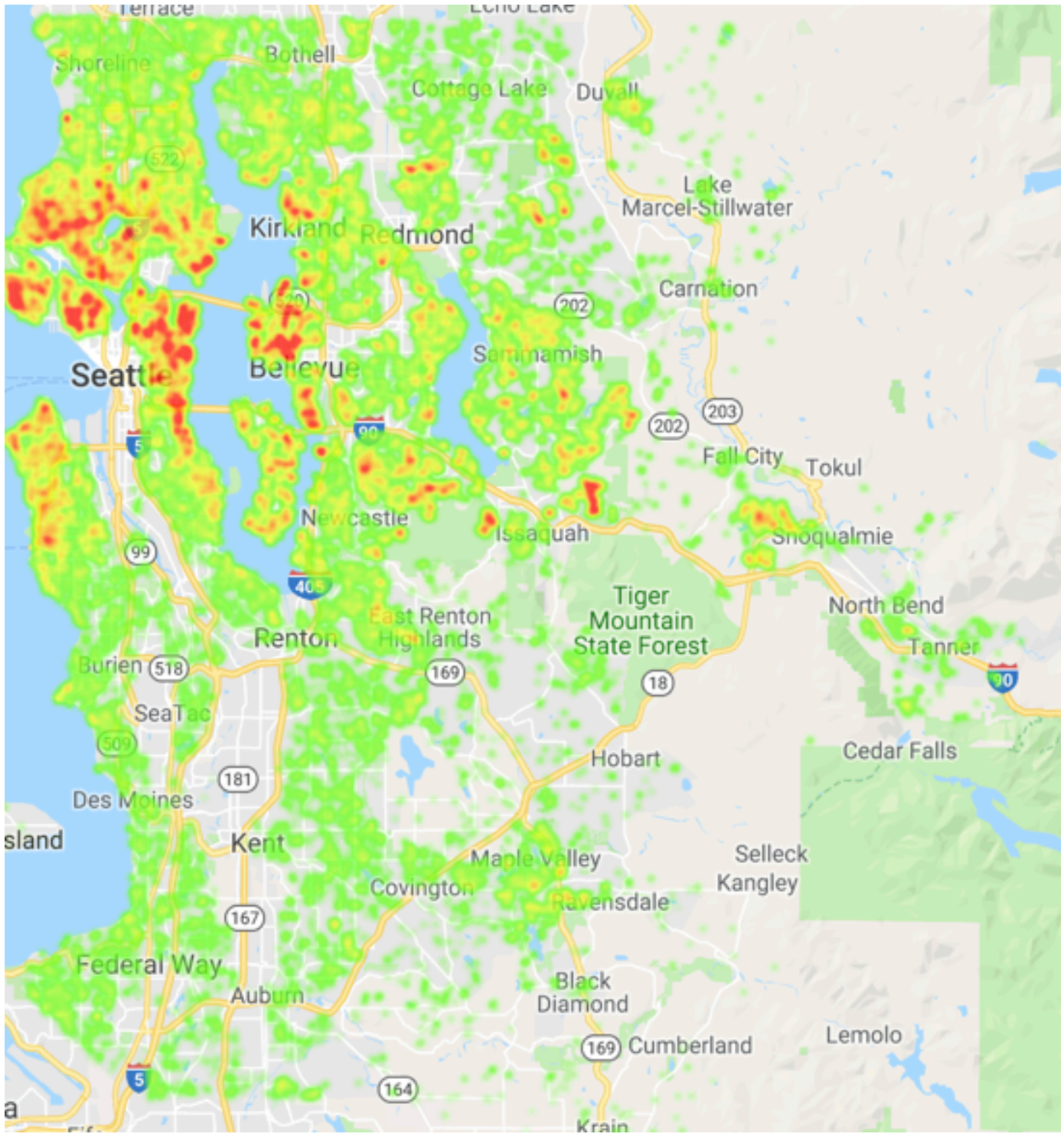
Outline

- **EDA**
  - Price distribution by location
  - Log transformation of key features and target
  - Multicollinearity
- **Model iterations**
  - Simple linear regression based on house size
  - Multiple linear regression models

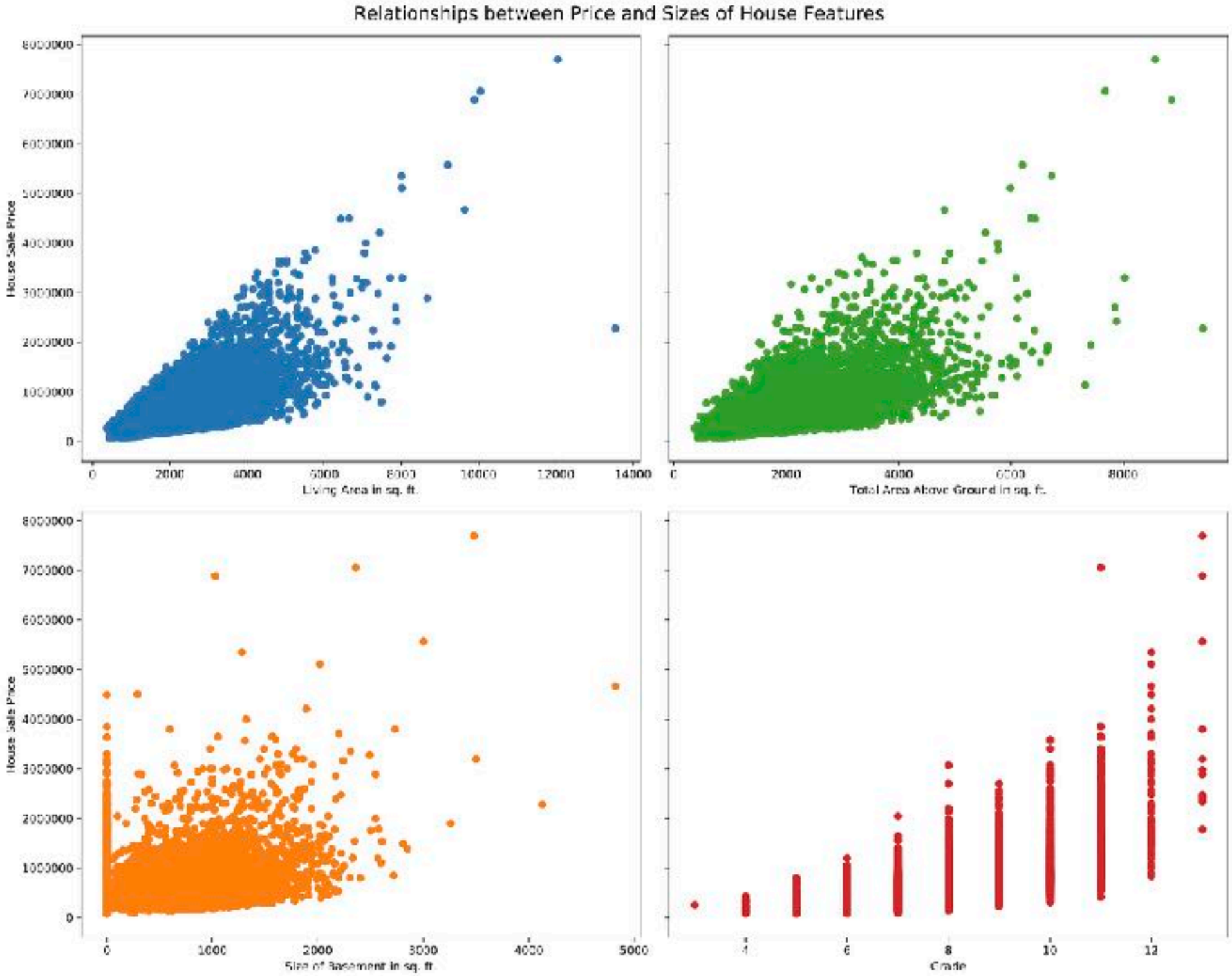
No obvious relationship between year house was built and sale price



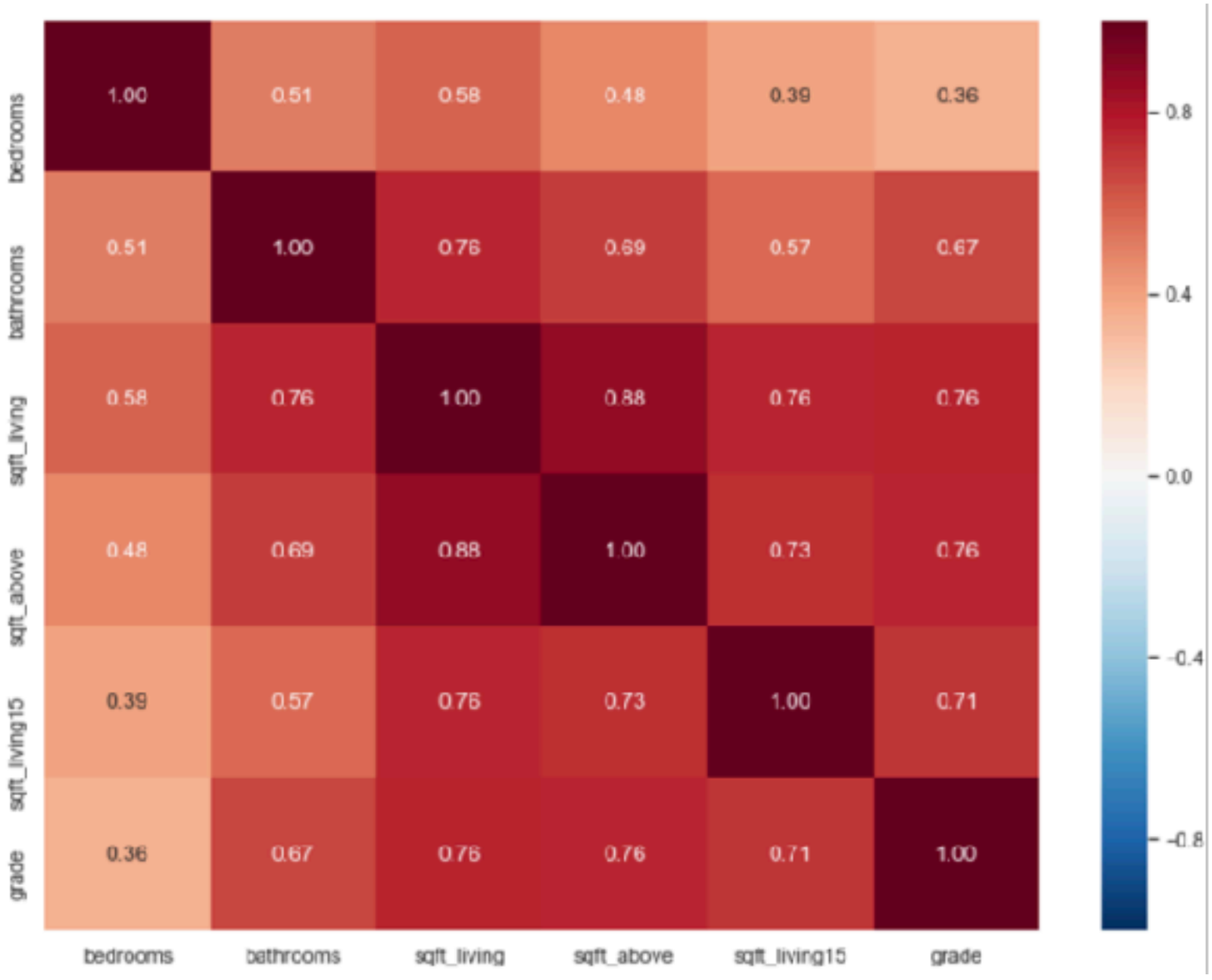
House price appears to be influenced by latitude and longitude



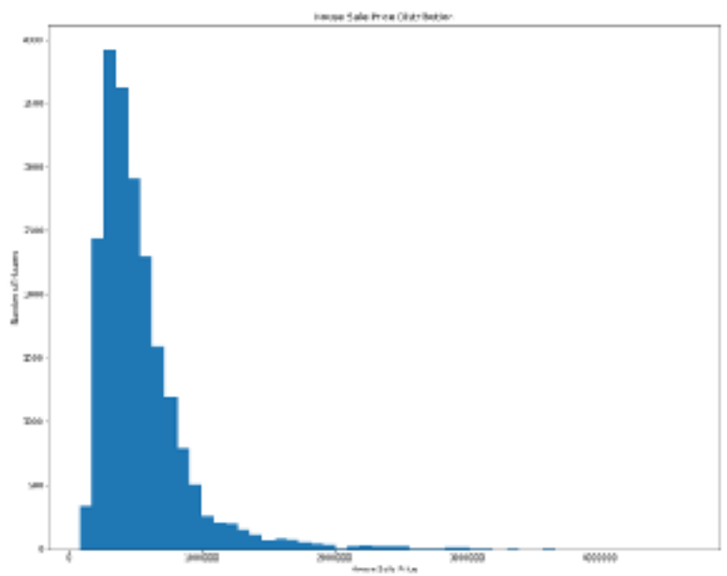
House sale prices in King County, WA correlate with size of living area

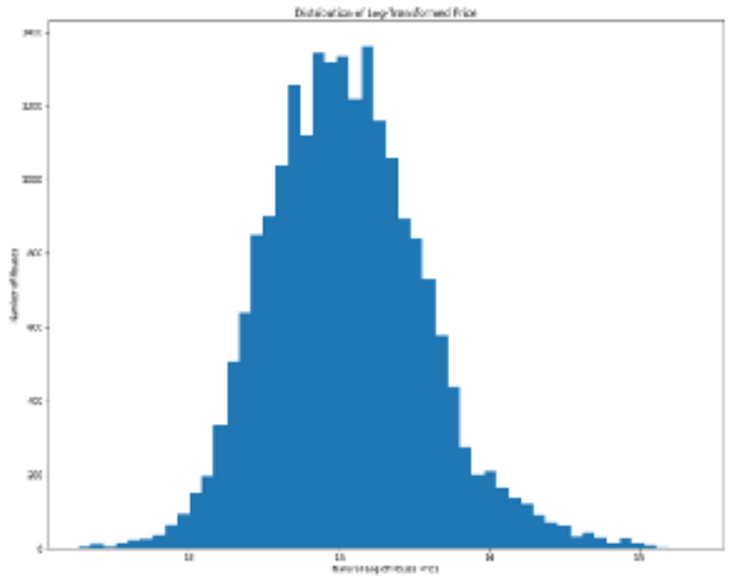


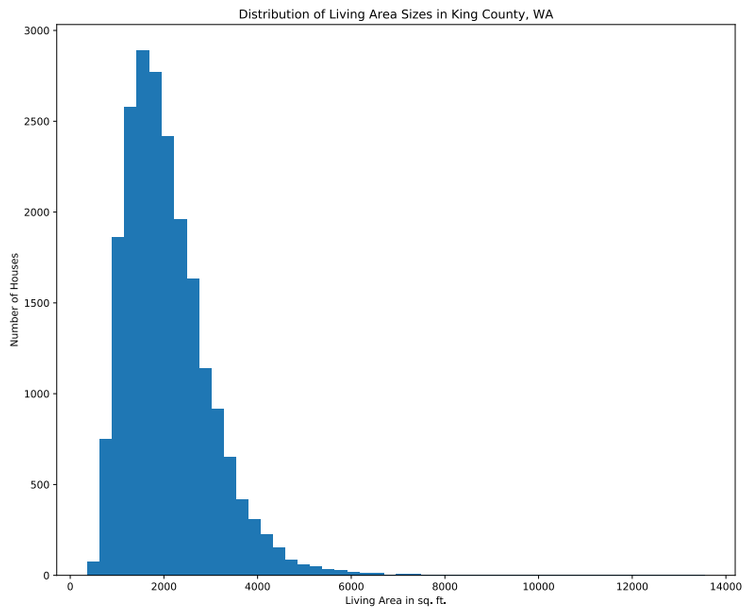
House features are highly correlated with one another

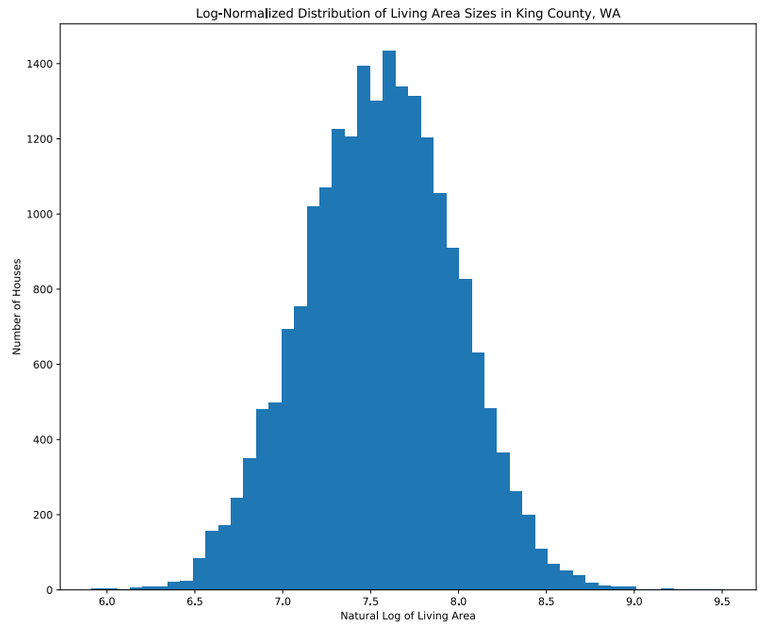


Log transformations for a more normal distribution

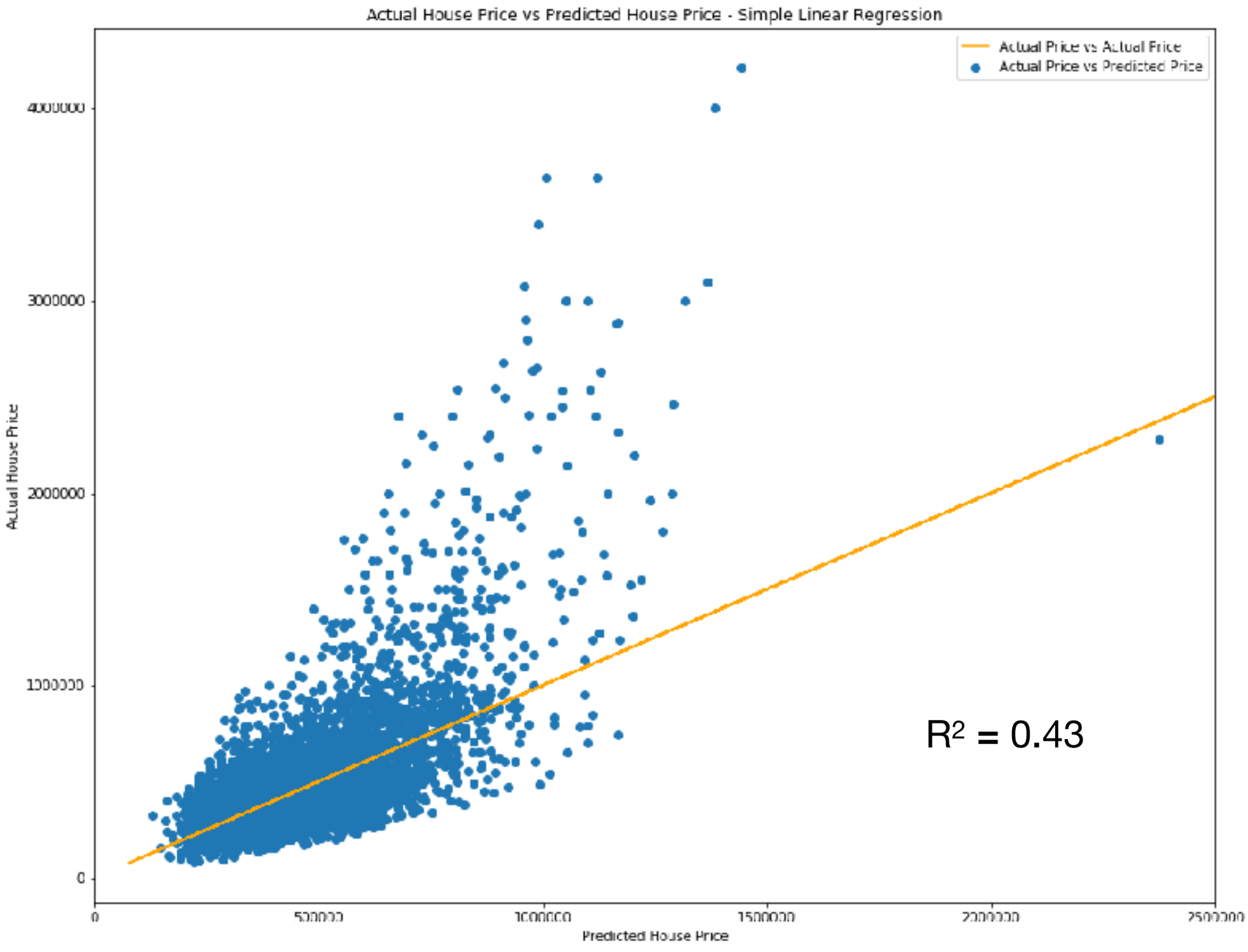




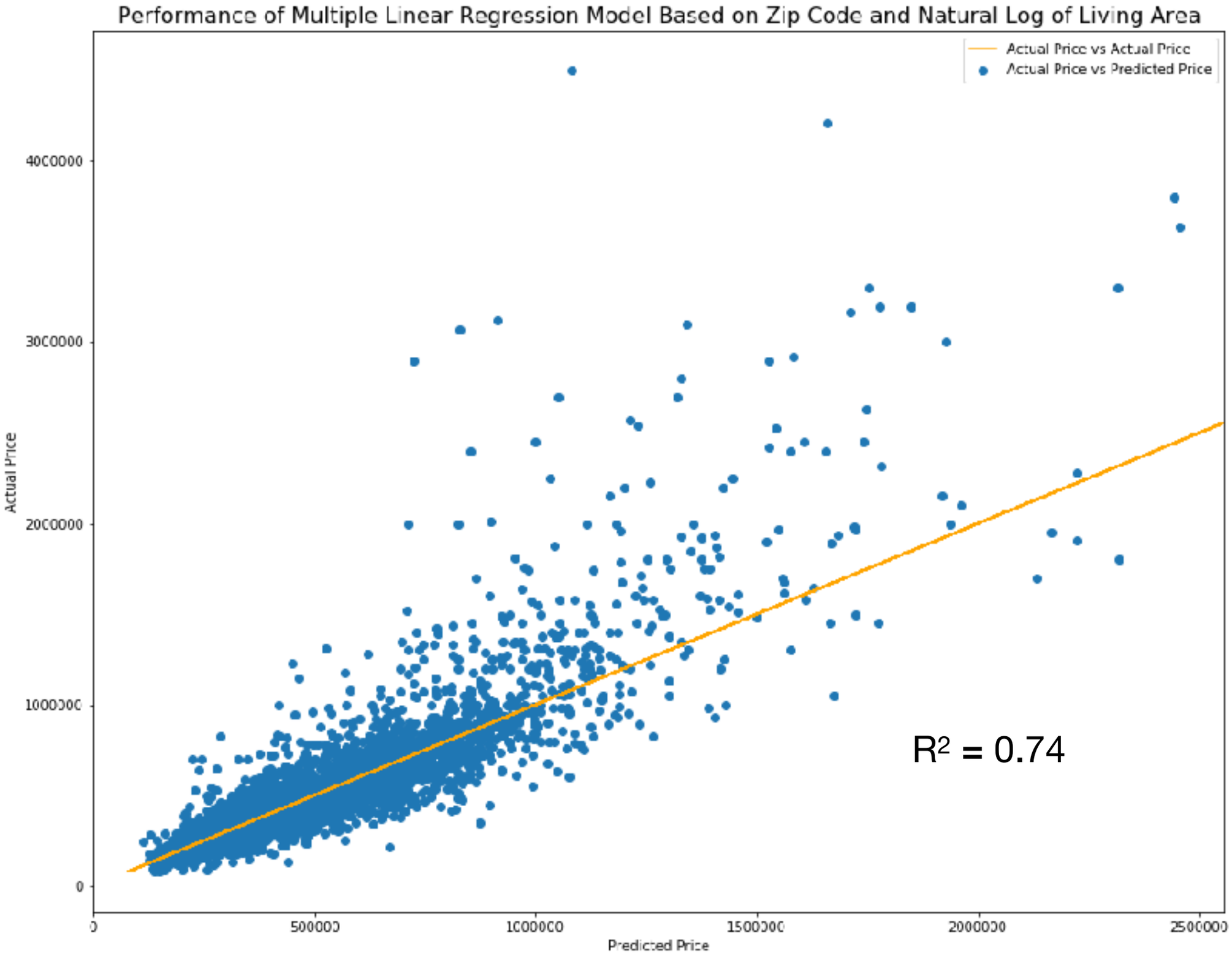




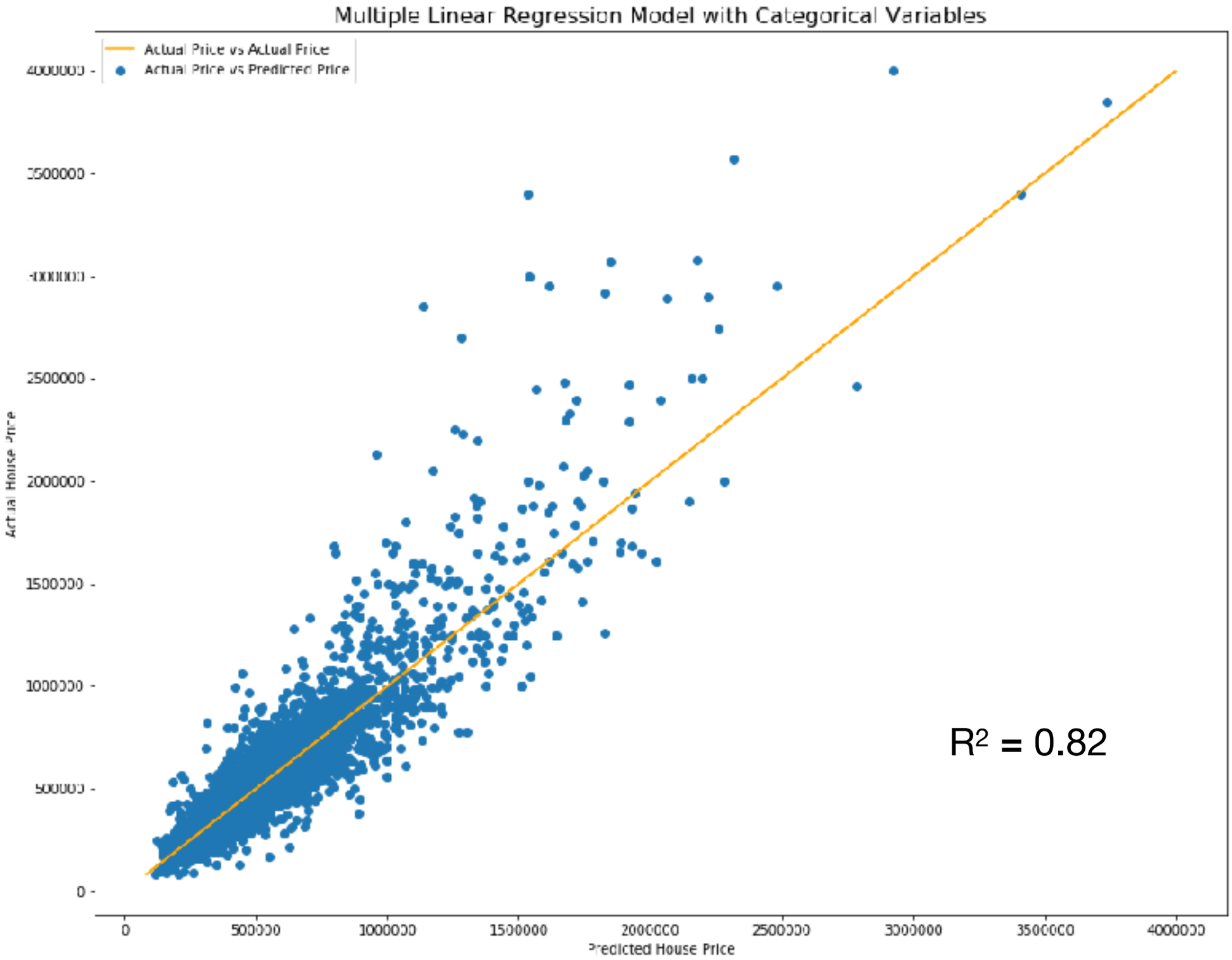
Simple linear regression model performs poorly



Grouping records by zip code improves model performance



Including categorical variables further improves model performance



Conclusions

- Price depends most heavily on size of living area and zip code

- Adding latitude, # views, whether a property is on the waterfront, and whether a house was renovated boosts model performance

- Our model struggles to predict the prices of the most extremely high-priced houses

Future Directions

- Model could be further improved by:
  - Feature engineering
  - Standardizing price and/or sq. ft. living area in a different way
  - Principal component analysis