The goal of my project is to predict the Seattle area's house market based on the existing features of a property. Seattle housing is a seller's market, the median day to sell is eight. This model would help buyers to have an understanding of how much a property would cost and help them make a rational offering.

Data: I scraped the redfin website's search result using Selenium to collect the four search results from four different districts. The properties are filtered for single family houses, townhouses or condos only. All the sales are finalized within the last six month.

I take out one feature: price per sqft since this element is calculated after sale and we usually won't have this information beforehand. Another part of data I took out is any property without a location.

Design and algorithm: The model starts with just numeric inputs, fitting into a linear regression model. The training data set has a R square of 63.6% while the validation data set is 63.3%. Not a great start but at least it's not overfitting.

To improve the fit of the model, I added an interaction term bed/bath. The feature increased the fit by 1%. I clean up the detailed location data to have no bucket less than 10 properties. After adding location as dummy variables, the fit improved to 68.6%, the difference between training and validation data sets is still small. To further improve the model, I added a deviation feature that describes how big a property is relative to the others properties within the same region. The reason is Seattle's eastside development model is very different from downtown. East side usually has larger houses built in the 90s, while downtown has new buildings with much smaller square footage. This feature improved the fit to 73.7%. I decided to take out on_redfin_days since this is another feature we won't know before sale, the fit dropped to 73% and is overfit. To reduce the overfit, I fit the model to Lasso and Ridge regression. They both reduced the overfit and Lasso has a slightly better fit.

Tools: Python, Selenium

Communication: Github