Linear Regression Project

By David Jia

First Look

Initial Questions about data:

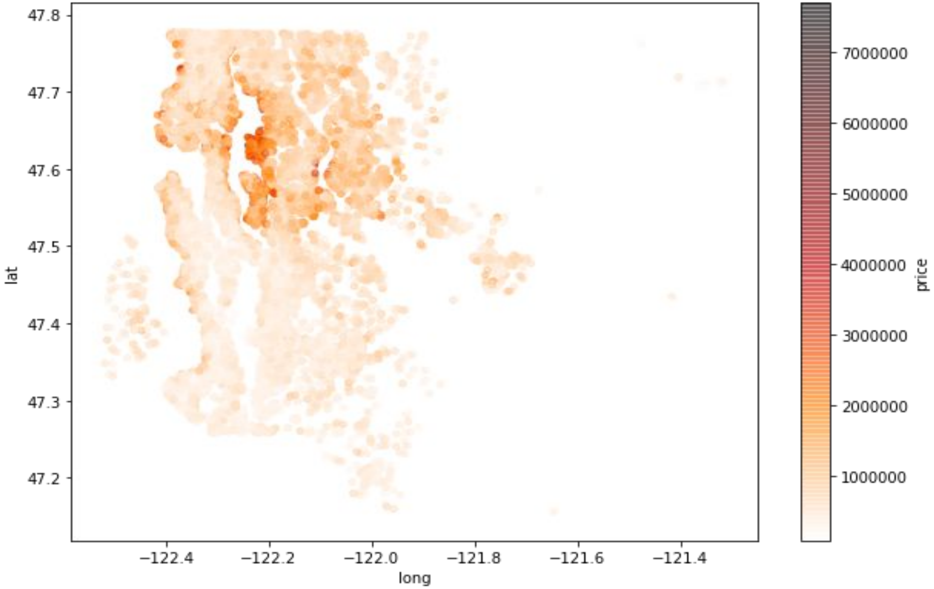
1. Is the data complete, what are the relevant variables in predicting price?

2. Does location matter significantly with price and other variables?

3. Does the year of renovation affect housing value?

4. Does a waterfront affect the price value?

Question 1. Exploring whether there is a hard correlation of latitude and longitude to price



-definite correlation in location with price

-clusters of high-worth houses grouped in certain areas

-seem to cluster near waterfronts

Question 2: Is a more recent renovation better for housing price?

Plot year renovated against price to see if more recent renovations lead to increased housing prices

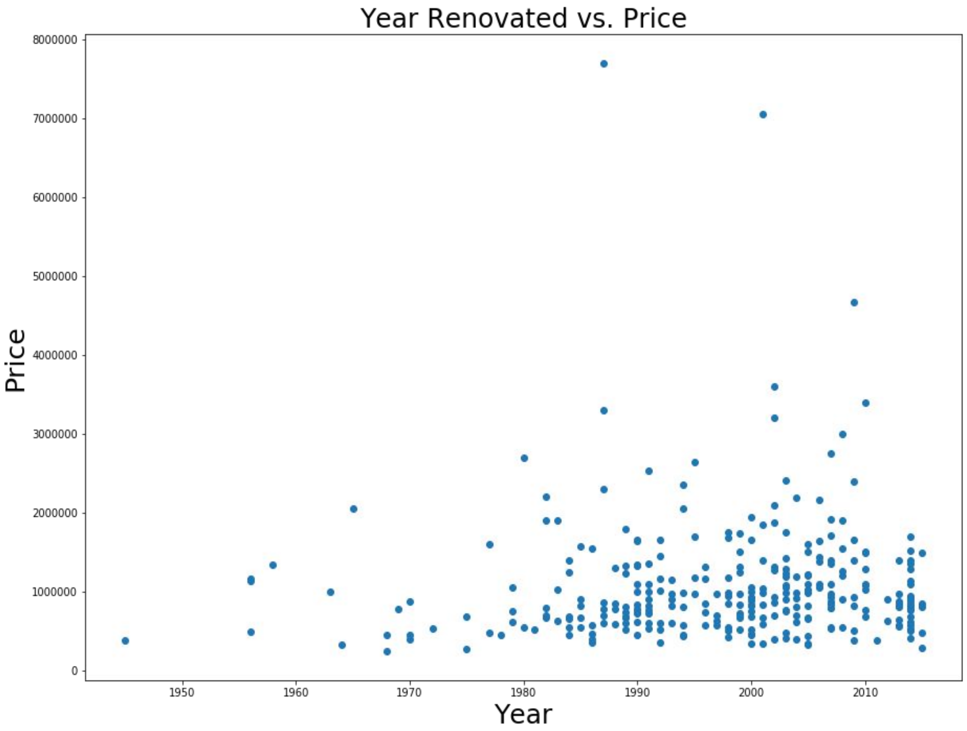
1. not the best data, missing a lot of values
2. not the best correlation either when ran in OLS

   R-squared = 0.012

   p=0.005

Probably little correlation

Final Answer: No, it does not.



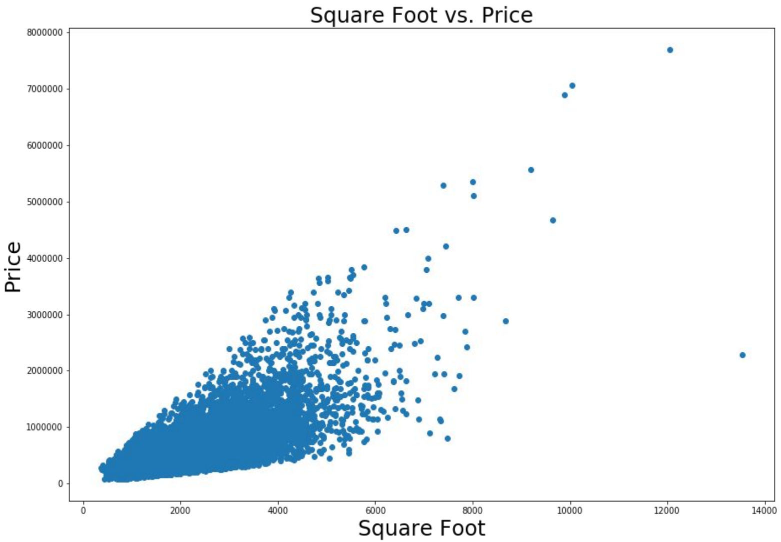
Question 3:
What are the most important variables in determining housing price?

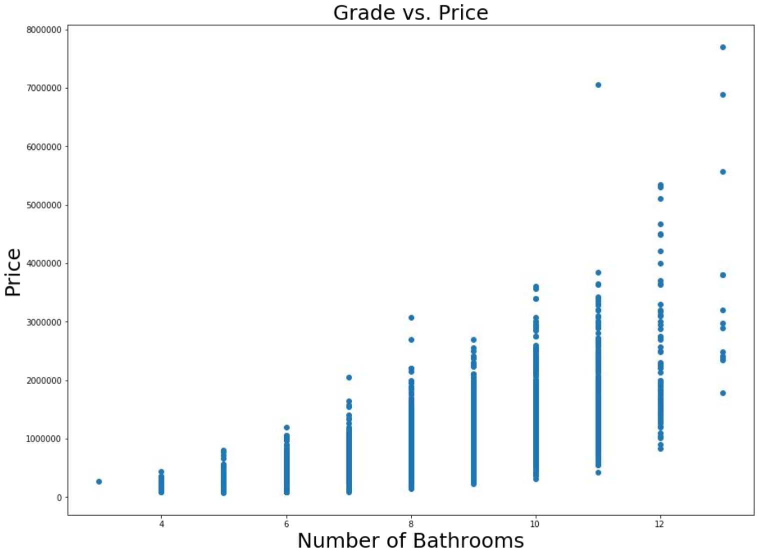
Approach:

1. Test relevant variables with OLS regression
2. Test for normal distribution
3. Look for multicollinearity
4. Feature engineer to reduce multicollinearity
5. Cross validate final model

First to study potential correlations with price, plotted every variable to price in a scatter

Only a couple showed good correlation: bathrooms, grade, and square feet measurements of the house





Related Variables

1. Square Foot of House
2. # of Bathrooms
3. Footage of House
4. Grade
5. Square foot besides basement
6. Square footage of nearest neighbors

After running regression tests, there seems to show a high amount of multicollinearity.

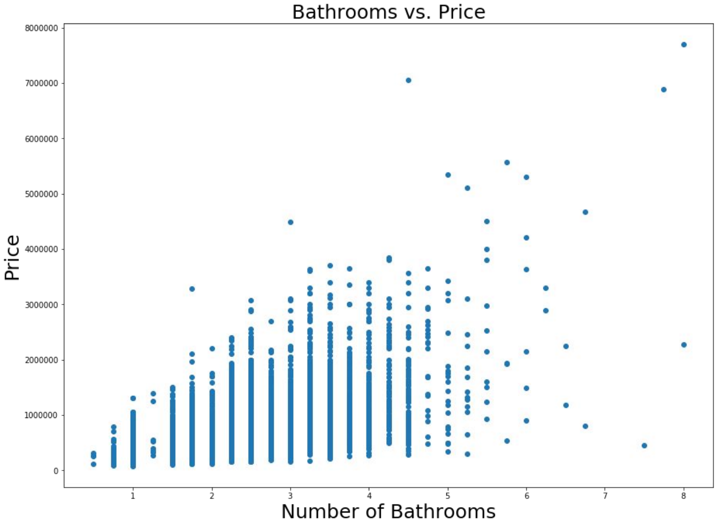
Found multicollinearity between:

1. Bathrooms and grade
2. Square foot of house, square foot of neighbors, square foot of house - basement

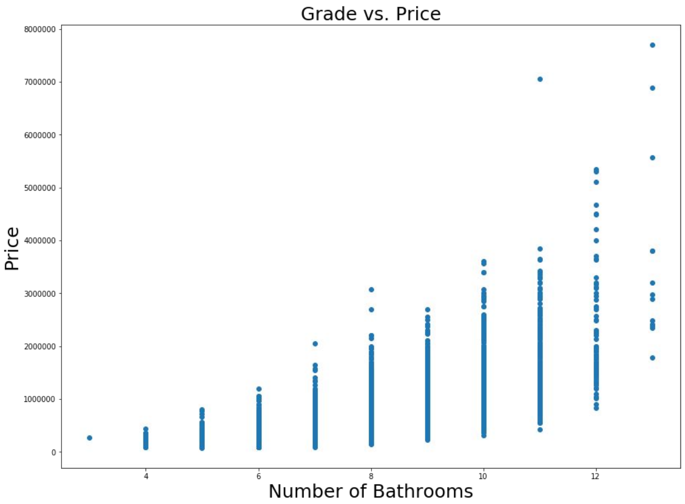
All these observations make intuitive sense.

Initially OLS Regression test shows:

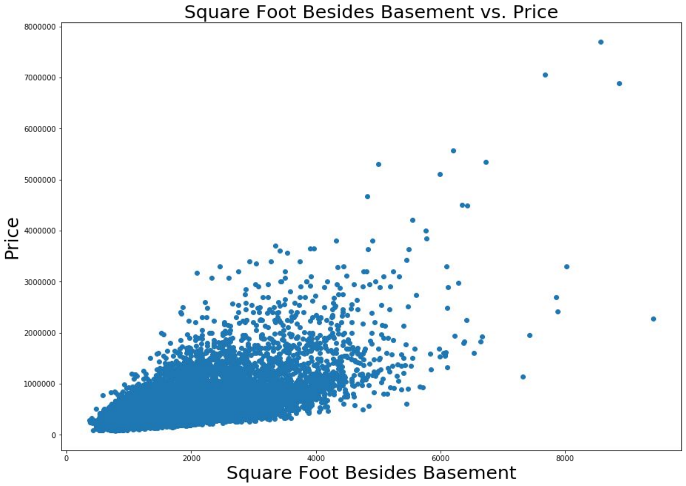
1. Decent R-squared: 0.504
2. Extremely high conditional number (multicollinear)



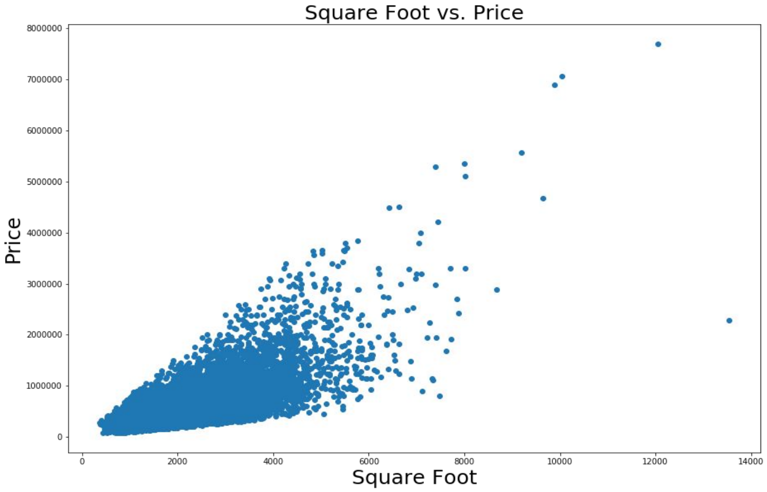
Bathrooms vs. Price



Grade vs. Price



Square Foot Besides Basement vs. Price



Square Foot vs. Price

Feature Engineering

Initial Steps:

1. Standardized the data for all relevant variables
2. Tried removing outliers but ended up keeping them

Combining Data:

1. Combined the datasets while averaging out for number of variables
    2. Resulting Data used in OLS regression testing

Final Model After Cross Validation

R-squared = 0.487

Conditional Number = 3.35

P values are appropriate (<0.005)

Final formula:

0.48x + 0.301y = Standardized Price

x=standardized square feet

y=standardized bathrooms + grade

Drawbacks include:

1. only King County homes
2. Not best R-squared
3. other significant variables unaccounted for