Module 2 Final Project

King County Housing Data

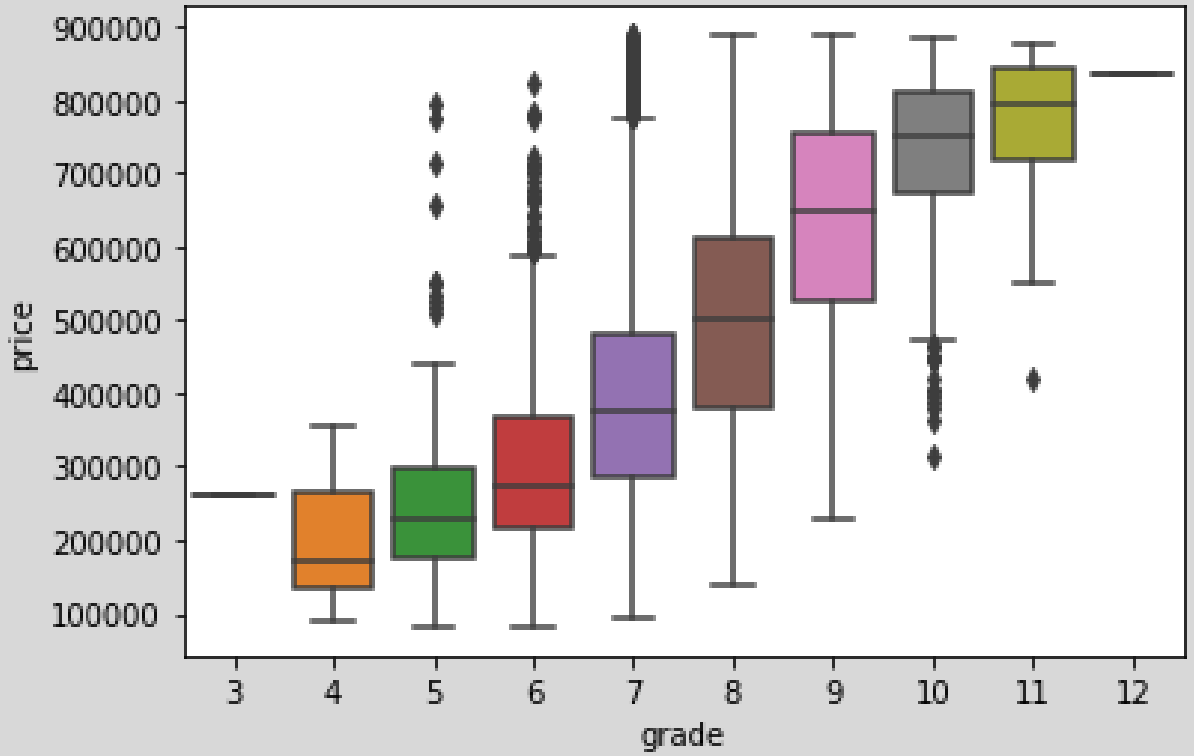
April 2021

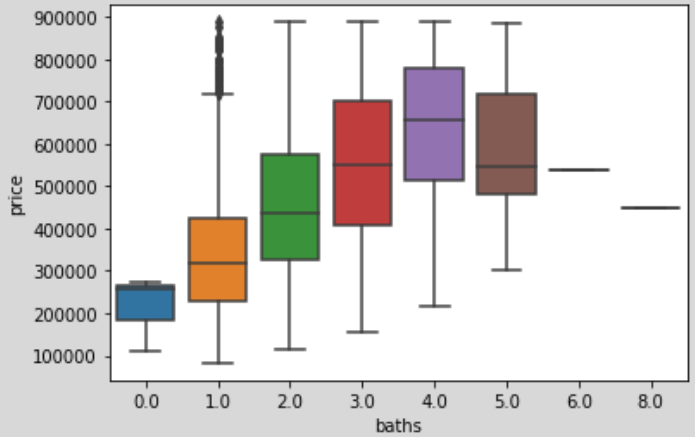
Lyndsay Maggin

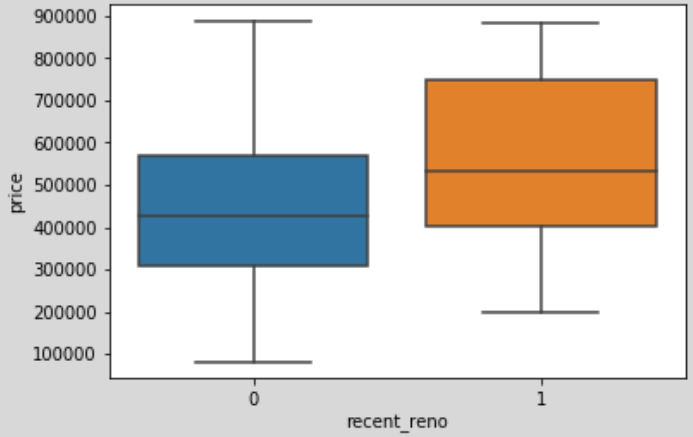
Scenario

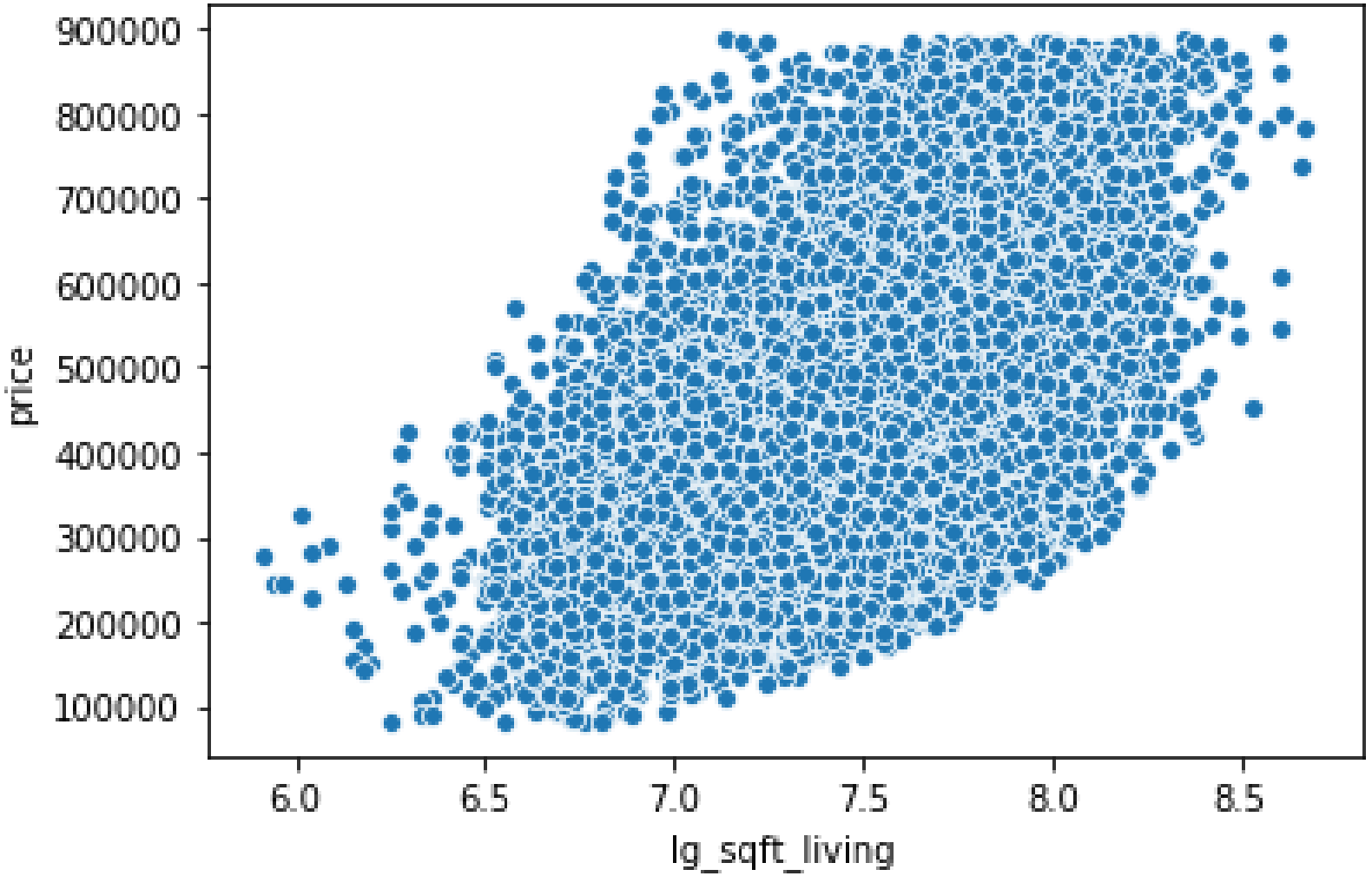
- It is 2016 and I am a realtor in the King County Washington area working with a new family to find their forever home. Using the 2014/15 King County Home Sales dataset, with my newly learned data science tools, I can predict the proper value of a property of interest and therefore can assist the family in entering a proper bid

Getting to Know the Data









Data Standardization

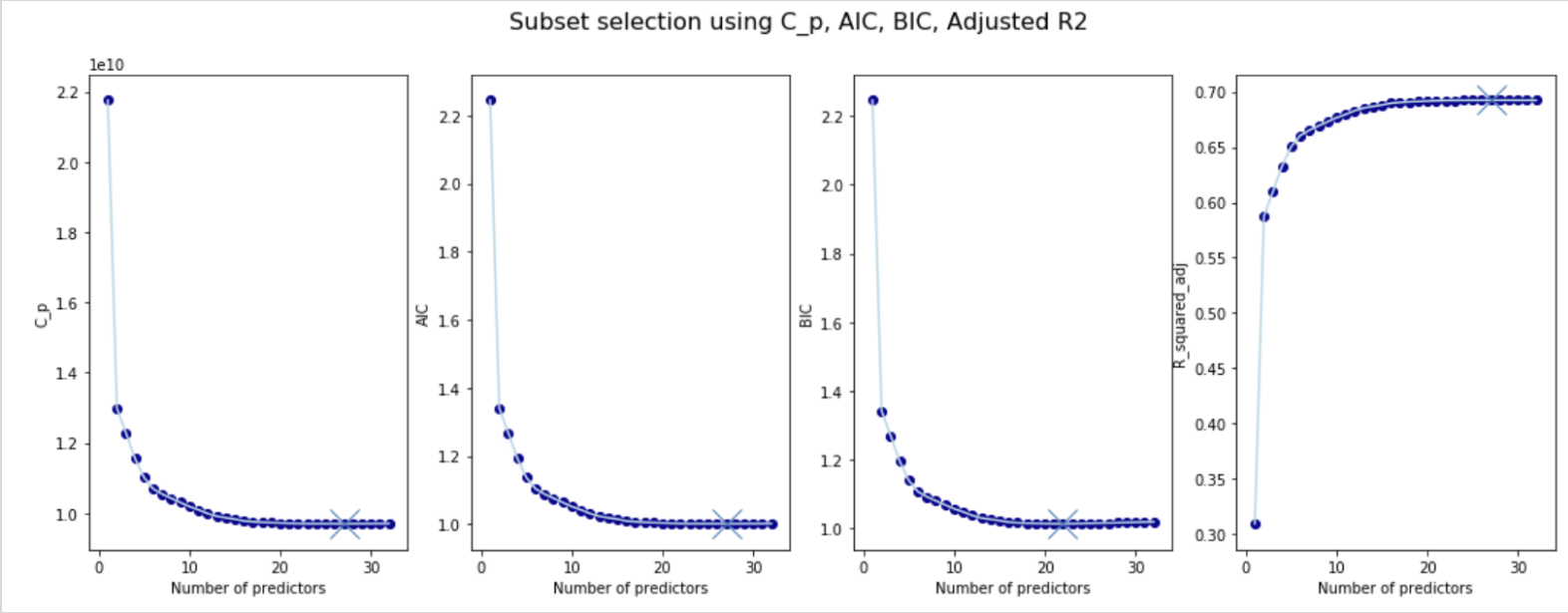
In order to improve normality of continuous variables, I logged and standardized.

Standardization also helps to reduce multicollinearity among predictors.

Correlation between sqft living and sqft lot was .14.

Choosing the Best Predictors

- Using Mallow's Cp, I determined 25 to be the best number of predictors for my model
- The more predictors you use, the less explanatory power you will have, I optimized the model using best subsets



Multivariable Regression

- Using K-Folds to train and test the model on 25 predictors, my model resulted in:
  - Rsquared: 69%
  - Mean Squared Error: $98,781
- Therefore I can assume that if I predict a housing price using my regression model, the price I predict could be off by about $98,781

Prediction

- My client is currently looking at the following home close to Medina:
  - 3 bed
  - 4 bath
  - 3,481 sqft livable space
  - Built in 2006
  - 16,988 sqft lot
  - gorgeous views of the water
  - highly rated
- The home is pricing for $899,900 today, is this a competitive price for them?
  - Using my model, I predicted a property value of $927,404.08
  - We can therefore assume this property is undervalued and my client should enter a bid
  - Adjusting for MSE the lowest bid I would suggest is ~$828,000 ($927,404.08-$98,781.00)