Predicting Housing Prices in Kings County, WA

Supervised learning capstone

By Sohaib Khuram

The Research Question

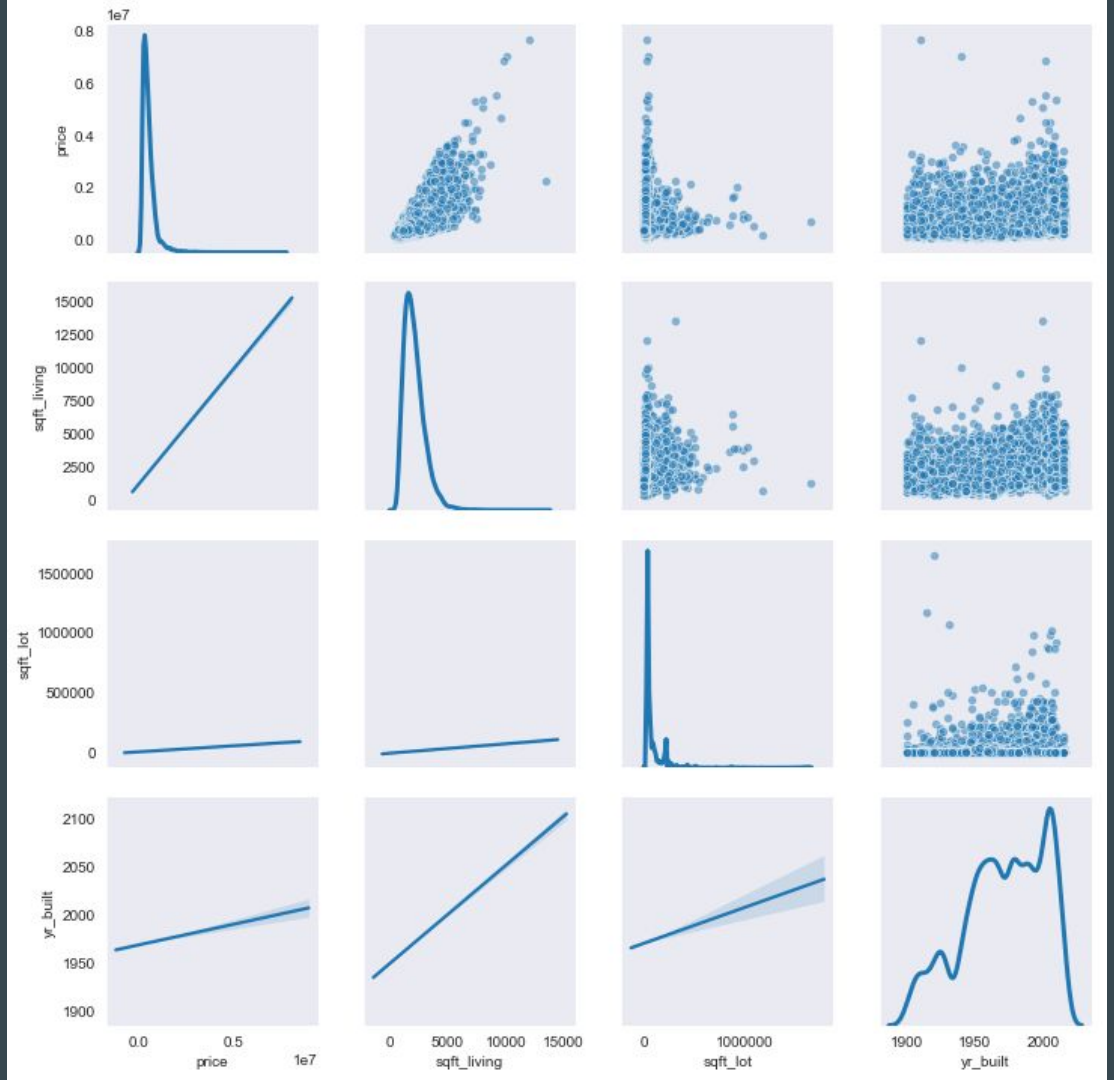
- How can real estate developers use knowledge about past sales to predict the sale of new houses?
- What features are important in determining price points?

The Dataset

- House sales prices from Kings County, which includes Seattle, in Washington state from May 2014 to May 2015
- Size: 21,613 observations
- Observations contain information about houses sold within the timeline along with info about property size, location, and features.
- Source: https://www.kaggle.com/harlfoxem/housesalesprediction

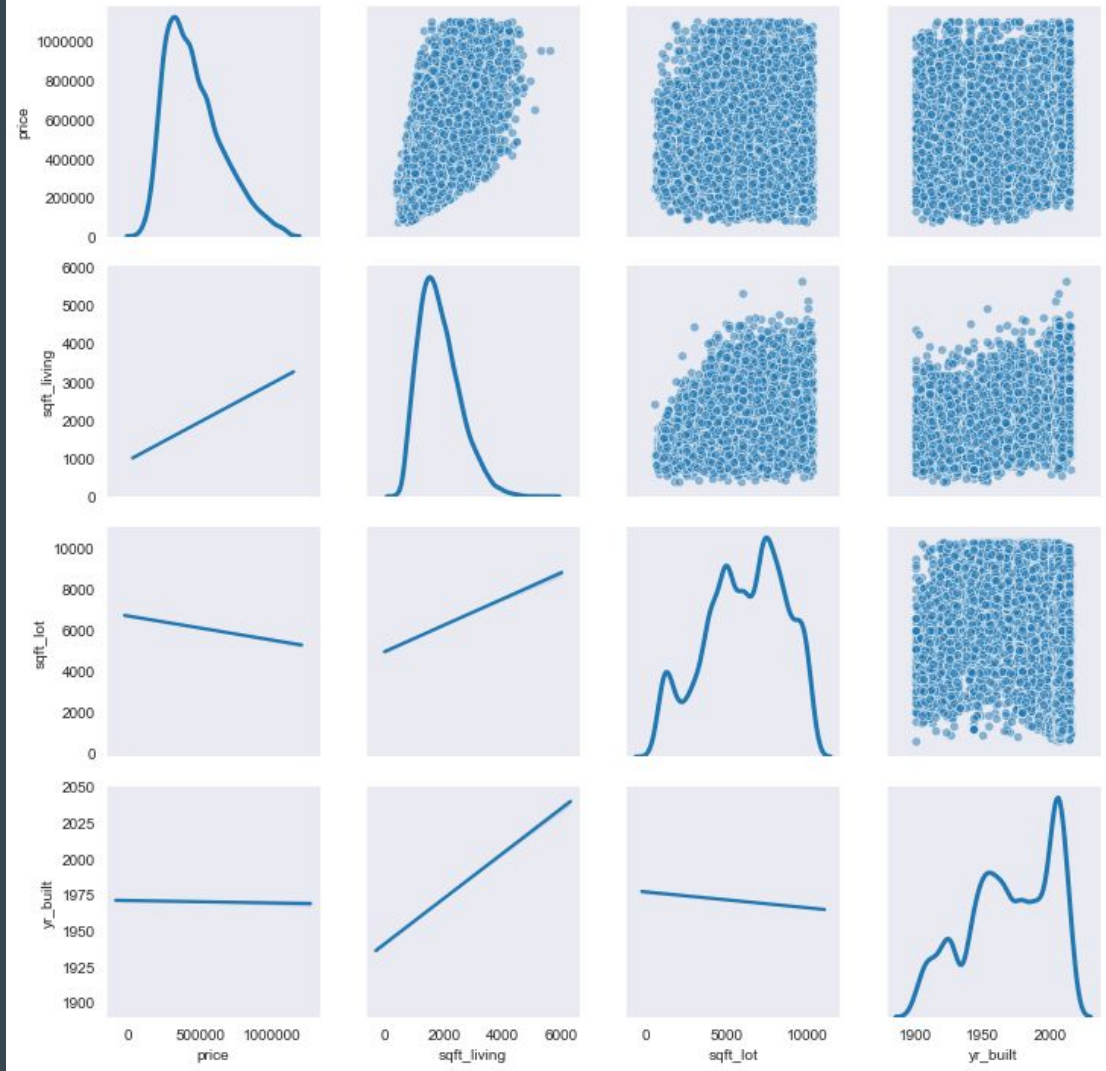
Inspecting the Data

- Making sure all data types are correctly cast and convert zipcode from int to nominal
- No null values in dataset
- Univariate analysis of target variable against predictors
- Create pair grids for continuous data and ordinal data to inspect distributions
    - Look for variables that have a linear relationship with the target variable
    - Year built seems like it doesn't have a linear relationship and a few other variables are redundant
- Shape of raw data: 21,613 observations


1e7
price
sqft_living
sqft_lot
yr_built
price
sqft_living
sqft_lot
yr_built

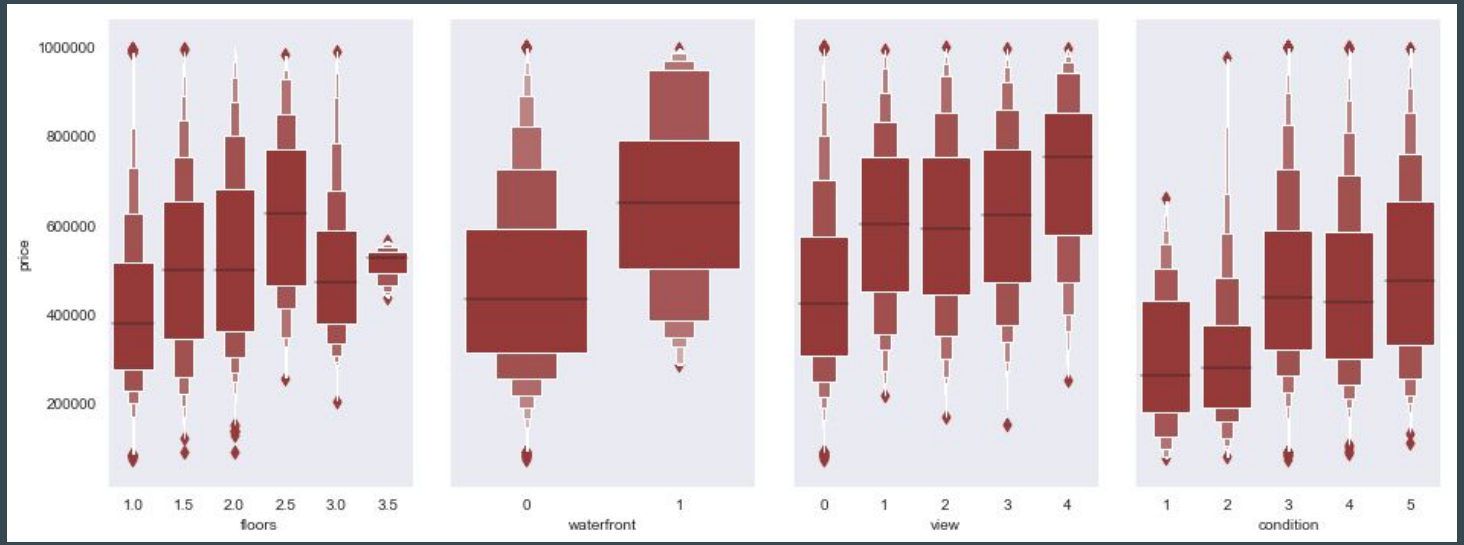
Outlier Removal

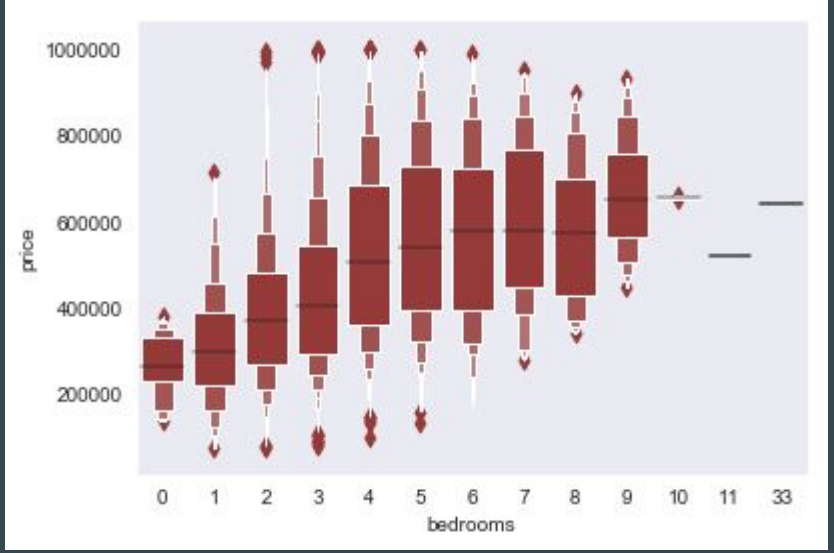
- The criteria for detecting outliers was based on whether entries were greater than 3 standard deviations away
- The criteria for sqft lot was more stringent as it's range is much larger
  - Values above the 75th quantile were discarded
- Removed outliers for sqft_lot and price
  - The higher end values for sqft lot that are left over are about the size of a city block (Manhattan)
  - Compared to properties that were a million sqft, these seem possible as houses


price
sqft_living
sqft_lot
yr_built
0
200000
400000
600000
800000
1000000
1000
2000
3000
4000
5000
6000
2000
4000
6000
8000
10000
1900
1925
1950
1975
2000
2025
2050
500000
1000000
5000
10000
1950
2000

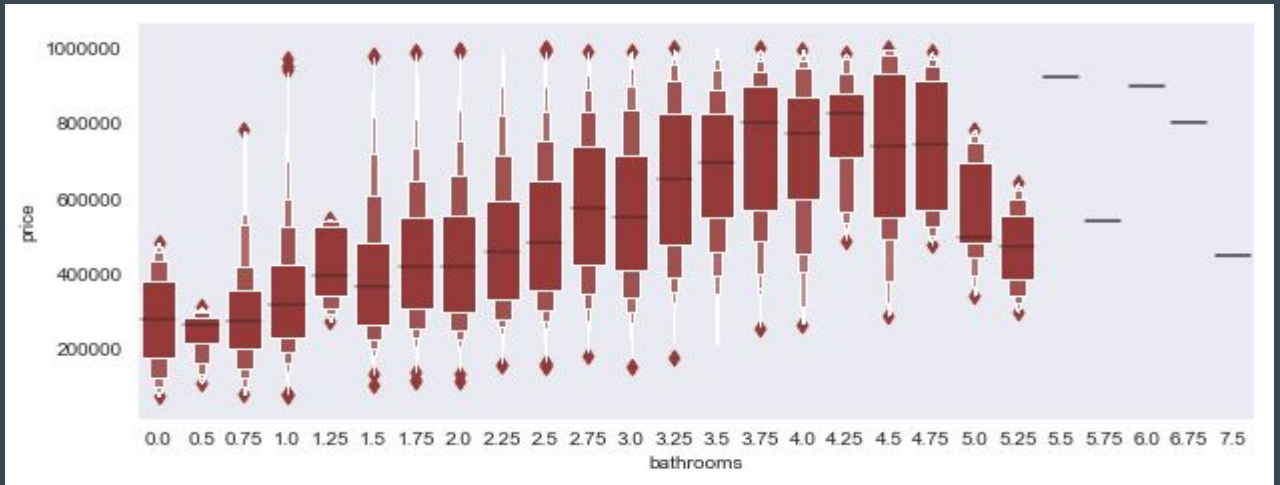
Ordinal/Interval Variables bivariate analysis

- Note about half floors: A floor is considered a half floor if the area of the upper floor is only 70% or less than the lower floor(s). Think sloped roofs and attics
  - Some notion of linearity between these variables and the price of the house





Distributions cont.



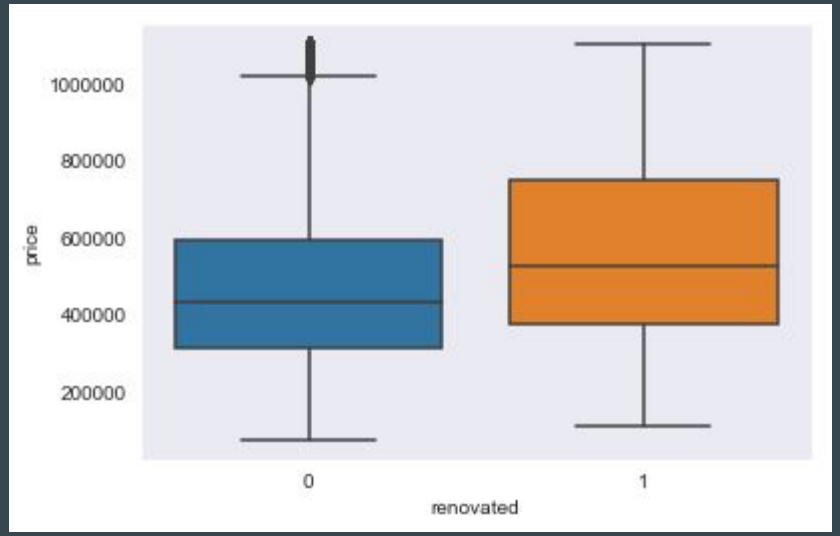
Outlier Removal

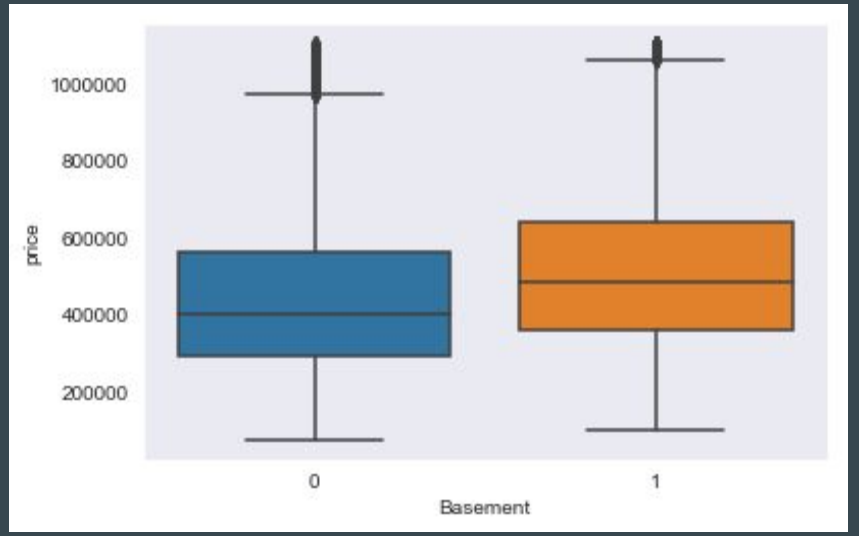
- Removed outliers that either didn't make sense or aren't part of the data needed to be analysed
  - A house with 33 bedrooms and 1.75 bathrooms. Also in only 1600 sqft so it must be one strange property or information was entered incorrectly
  - Properties with 0 bedrooms and 0 bathrooms

Data Preparation for Modeling

- Feature Engineering
  - Create binary features for whether a property has a basement if sqft of basement greater than 0
  - Create binary features for whether a property was renovated if yr_renovated greater than 0
  - Create the 70 zip-codes into dummy variables
  - Created log-lot to make for a more normal
- Drop redundant variables such as sqft_living 15 and sqft_lot15 to reduce noise in data
  - Also drop the variables used to engineer our earlier binary features to reduce collinearity
  - Drop lat and long as zipcode covers geography

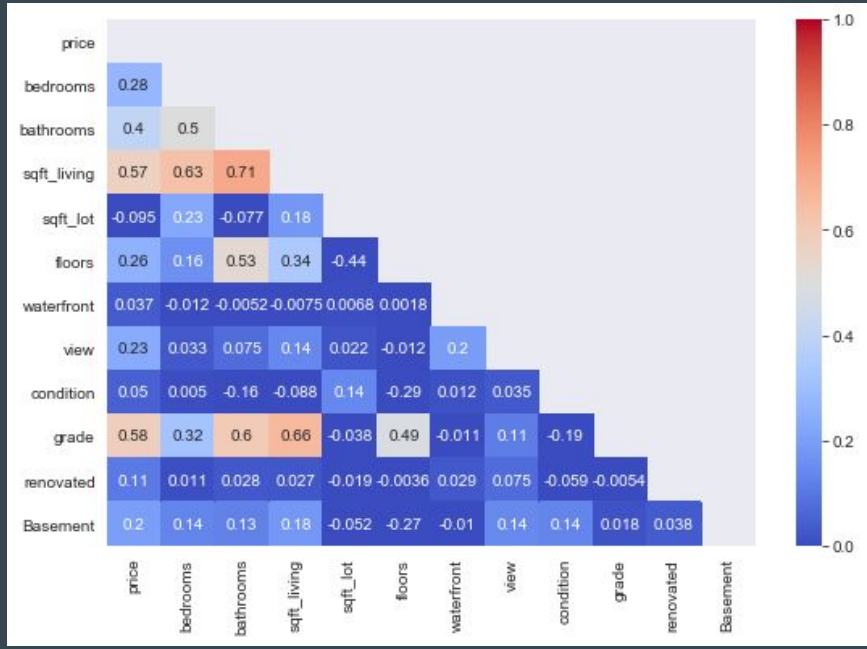
Distribution of Engineered Features





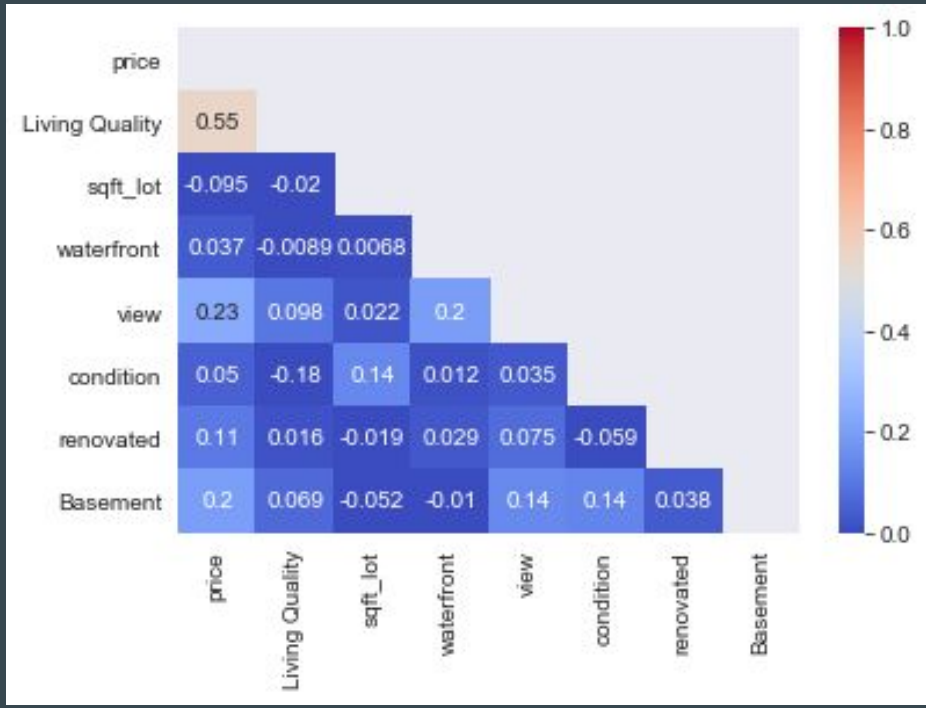
Collinearity Checks

- A few variables are correlated and represent similar/related ideas, so they are a good target for PCA
  - Bedrooms, Bathrooms, Grade, Sqft_living, Floors
- Use PCA to reduce linearity between the predictors



PCA

- PCA run on bedrooms, bathrooms, grade, sqft_living, and floors to create the feature called Living Quality
- Multicollinearity issues resolved



Scaling features to make a robust model

- With the exception of principal components and binary variables, all variables were normalized by their mean and standard deviation to account make the scale of all predictors the same
  - Used in Lasso and Ridge to improve regularization

Exploring Models

- Create a dataframe to hold results of multiple models to find the best one
- Based on the correlation matrix of the original variables, sqft_living was the most highly correlated with price, so it's a good candidate for a simple linear regression
- Make a multiple variable regression using all the basic features

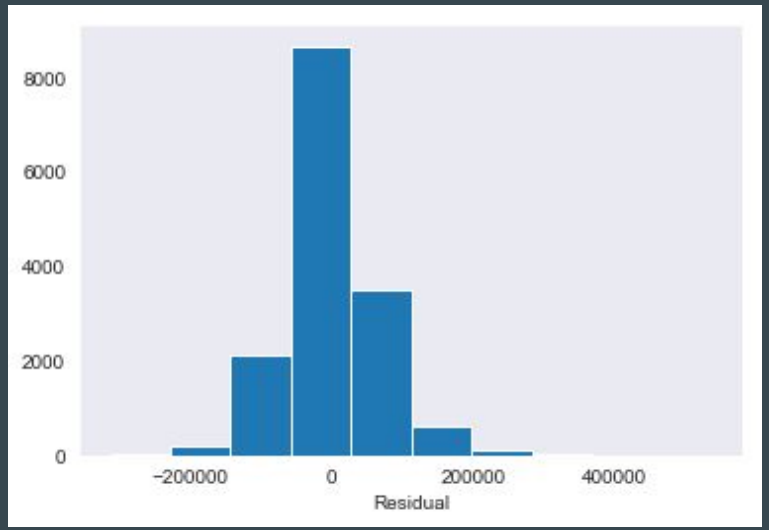
The Prime Candidates

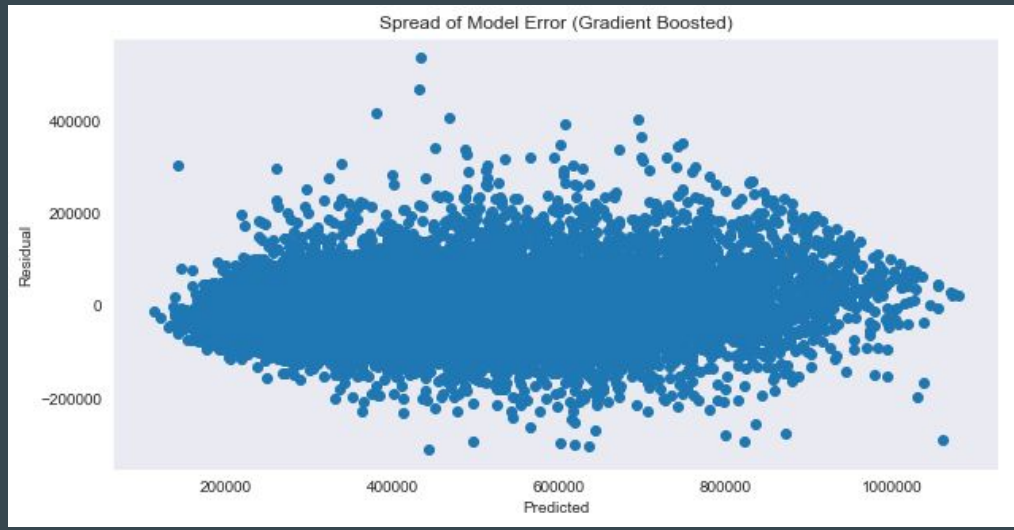
- The Gradient Boosted Regressor and lasso perform fairly well
  - Use lasso to understand weights of the features
- Use multiple regression after PCA, Ridge, and gradient boosted regressor as alternate models to consider

	Model	Mean Absolute Error (MAE)	CrossVal Score (avg)	CrossVal variance
0	Simple Linear Model	132910.652	0.32094	0.0290966
1	Multivariable Regression (PCA feats)	67876.164	0.778496	0.0293775
2	Ridge	65173.675	0.797973	0.0166751
3	Lasso	60098.064	0.822185	0.0129906
4	Multivariable Regression (non-PCA)	61512.172	0.822185	0.0129906
5	Gradient Boosted Regressor	55321.198	0.84108	0.0115533

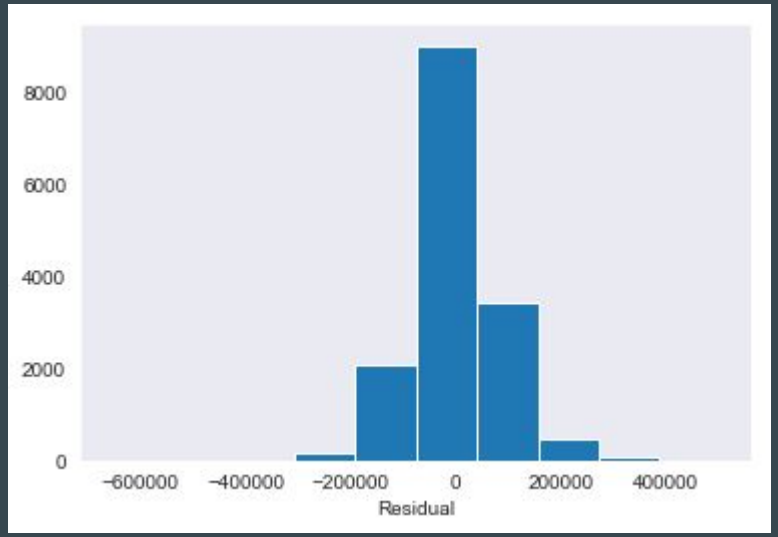
Final Model Choice

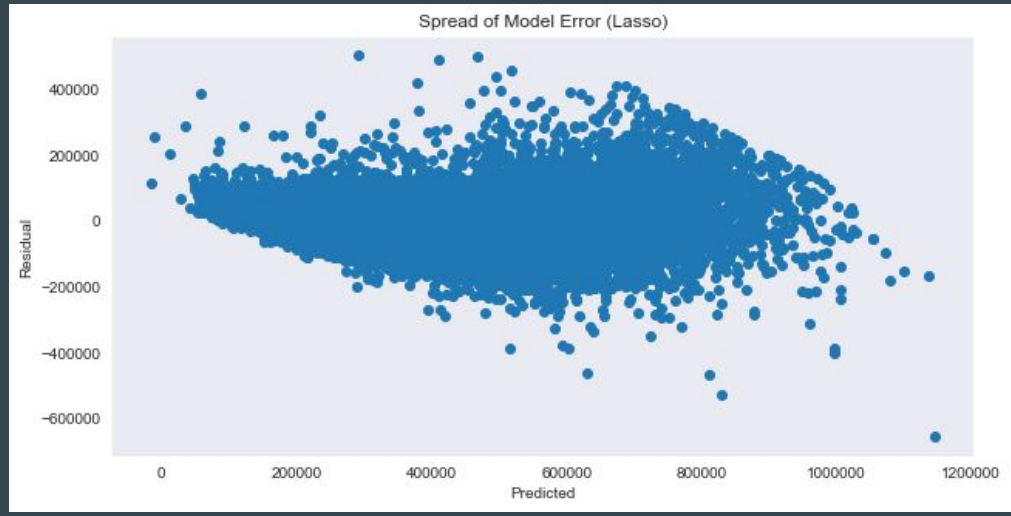
- The gradient boosted regressor is the best bet so far as it offers a good mix of being an efficient model, with high accuracy, and minimal variance in the error distribution





Lasso validation





Cross-validation results

Gradient-boosted Regressor:

```
array([0.83023424, 0.83146393, 0.85784605, 0.82394746, 0.83332932,
       0.84643326, 0.85200513, 0.85518488, 0.84858868, 0.83177159])
```

Lasso regression:

```
array([0.81709456, 0.80962471, 0.83608988, 0.80126905, 0.82402233,
       0.82827644, 0.83511423, 0.83945487, 0.82721336, 0.80369506])
```

Insights

- The gradient boosted model, as expected, has incredible performance, runs faster, and is more flexible than the lasso/multivariable regression model
- From lasso, factors such as bedrooms, bathrooms, floors (makeup of house) did not add as much value as the size of the property and which zip-code it was in.
    - A property in Seattle can be worth more than $25,000-$50,000 with the same property features as another house in a different part of King County
- Some factors such as bedrooms and floors had negative coefficients
    - This can be explained by understanding how changing those variables while keeping sqft_living constant would affect the house price

Insights cont.

- Proportionate increases in living quality (bedrooms, bathrooms, floors, grade, sqft of living space) lead to increases in prices more linearly
  - Making houses with 6+ bedrooms, 1.5 floor and 1 bathroom wouldn't make you more money than a house with 3 bedroom, 1 bathroom, 1 floor.
- Properties starting with zip-code 981 tended to be slightly more expensive than those in 980
  - The 981 zipcodes are are the north+west parts of King County (area includes Seattle) and 980 is the South+West parts
- Model seems to be better at estimating prices rather than predicting them

Limitations/Design choices

- Not tested on other counties, so model could only be generalizable to the King County area
- Specific property types not included such as properties with odd house qualities (33 beds 1 bath, 1 million sqft properties, no bathrooms)
- Only sales from one year, so estimations do not account for inflation
- One of the wealthiest cities is included in this county, which can cause price disparities between property types
    - A 1000 sqft property in Seattle would be much more expensive than a 2000 sqft property out in the suburbs or more rural areas.
- Model can be improved as our prediction values are not consistent across all ranges

The End