King County Real Estate Predictive Modeling
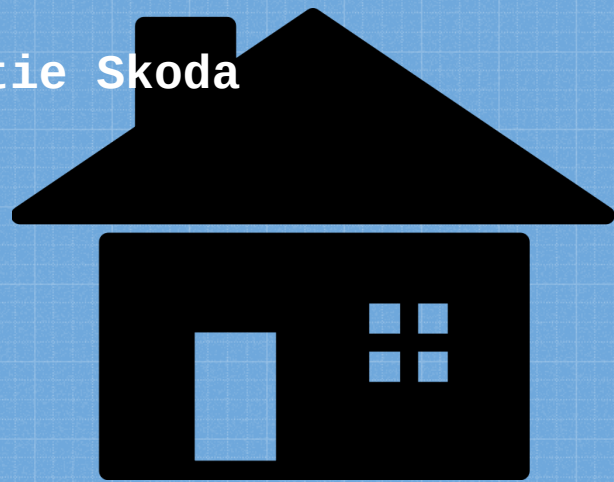
Colette Crowder, Joe Swing, Katie Skoda

Cleaning

- I replaced all '**?**' with NaN values

- Checked for null values and dropped the ones found
  - 18 of the 21 column did not have missing values
  - Year renovated had 17% values being missing
  - Waterfront had 11% values being missing
  - View had .2% values being missing
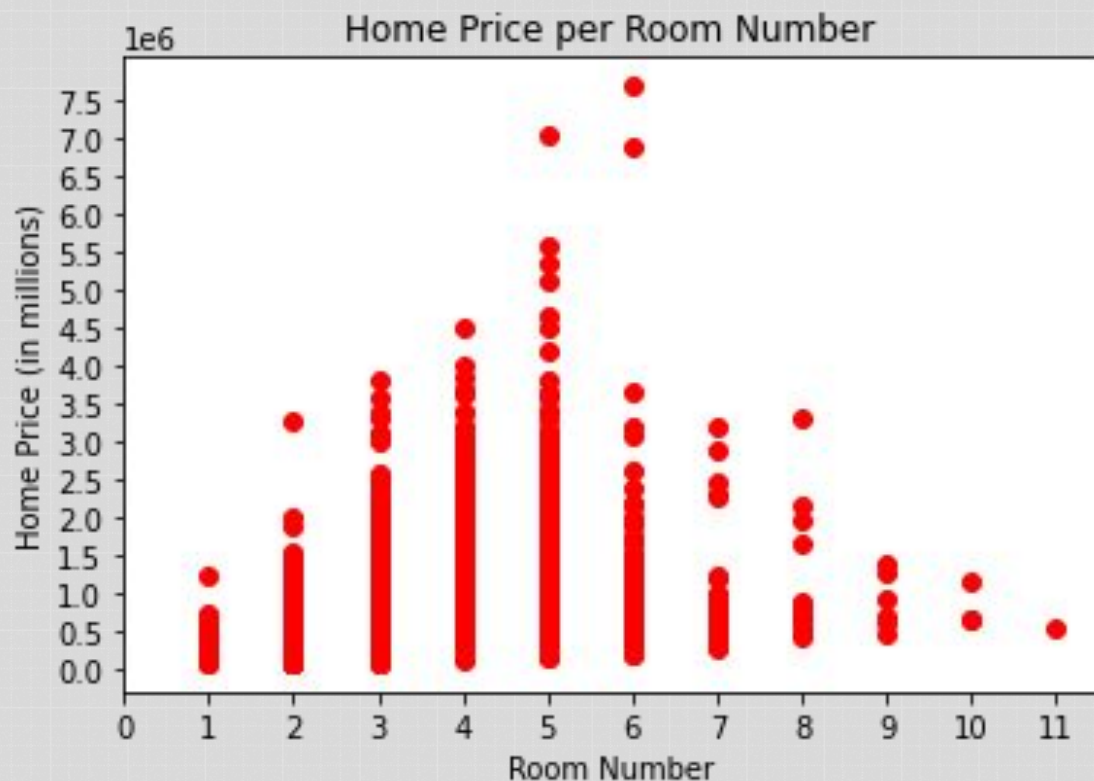
- Removed the entry that had 33 bedrooms

Correlations Between Variables of Interest

How Footage of Home Affects Price

Home Price per Room Number
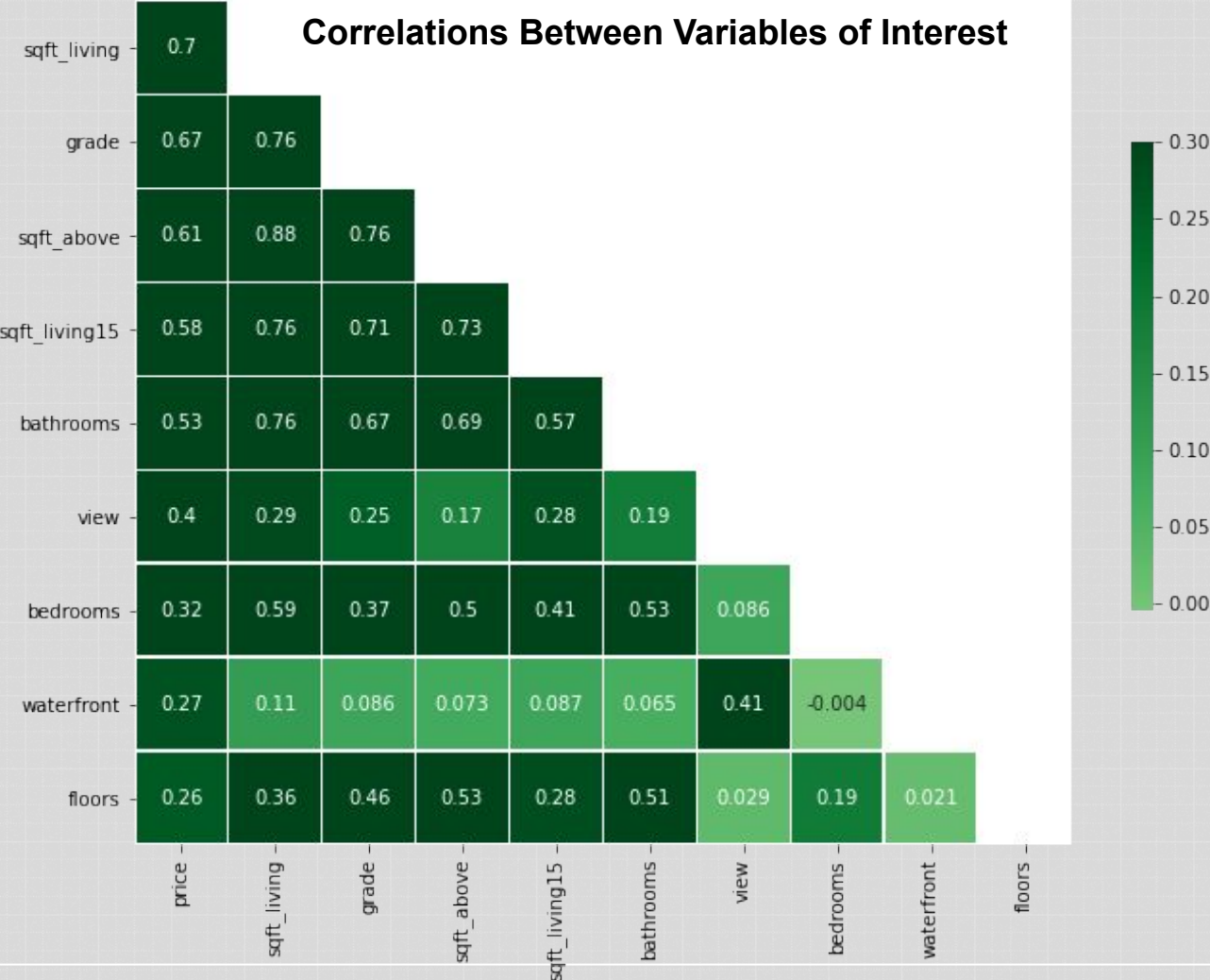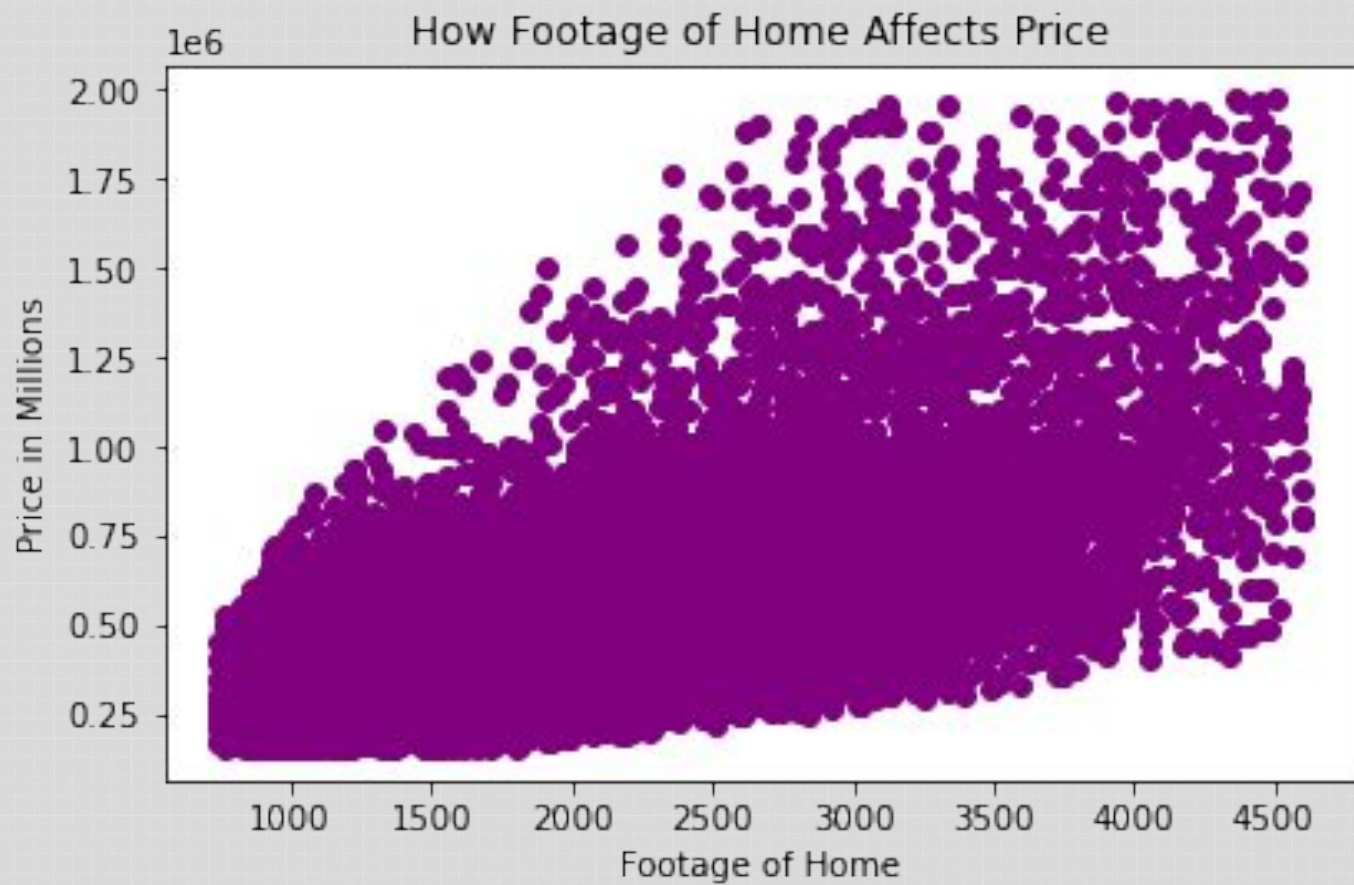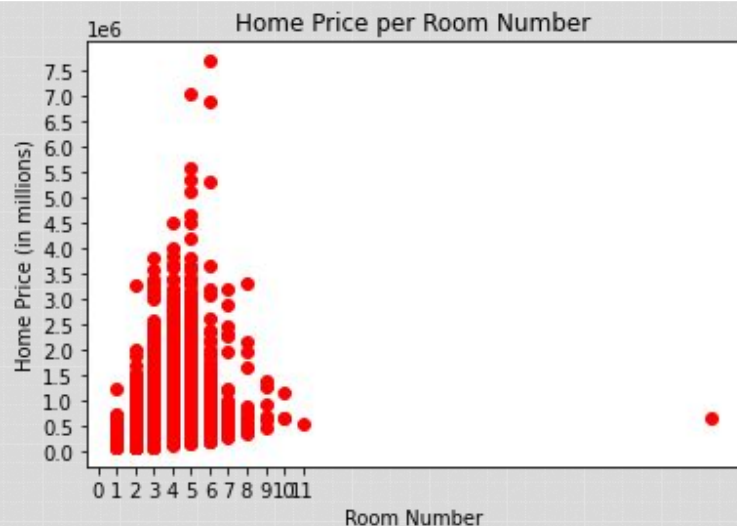
Dropped the outlier of 33 bedrooms for a better visualization and more precise data
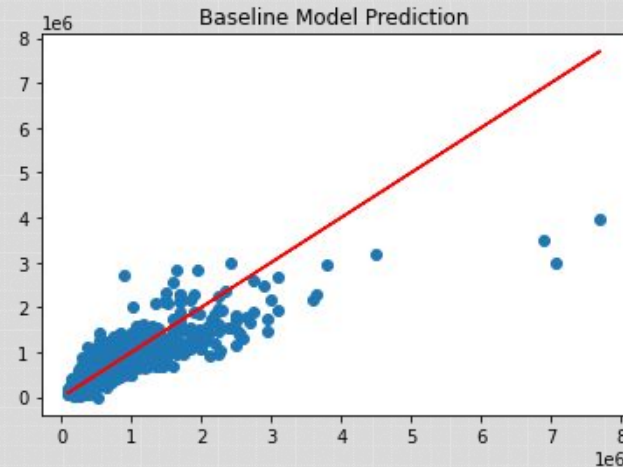
Baseline Model

- For the baseline model, we included every variable except for id.
- The scores we got for our testing and training data were:
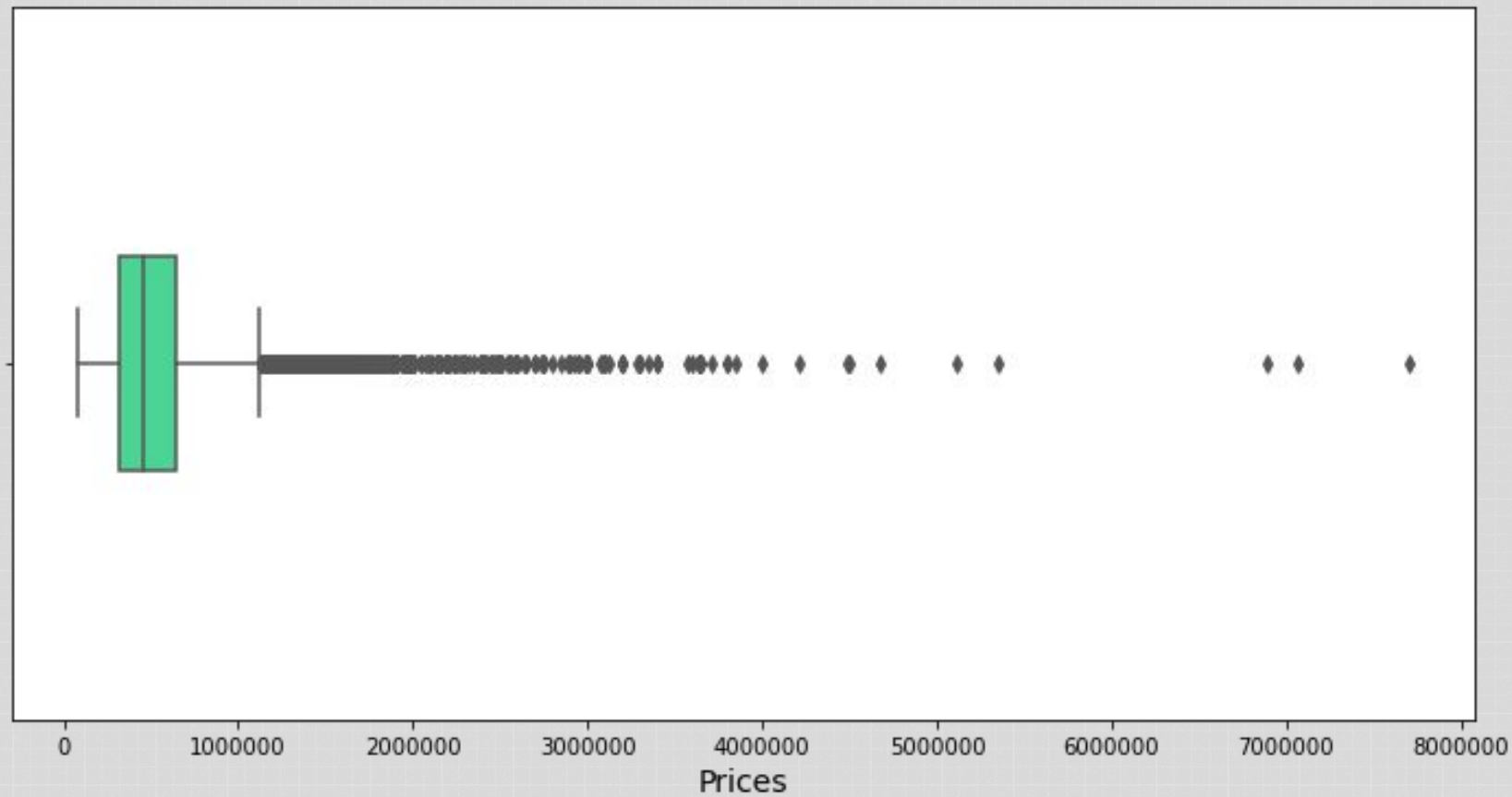
Training R^2: 0.7339

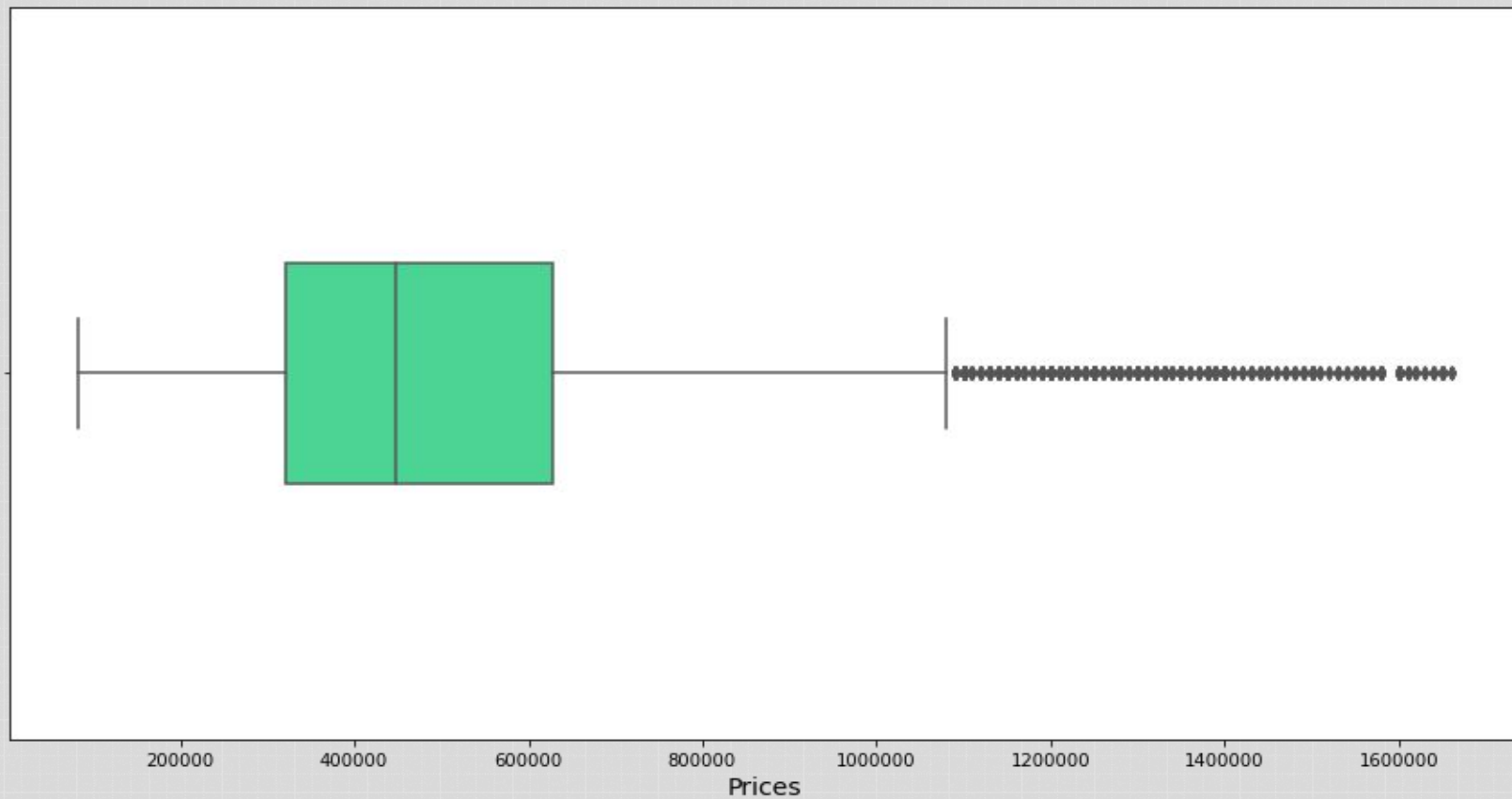**Testing R^2: 0.7184**

Training RMSE: 185625.8124

Testing RMSE: 217398.4034

Baseline Model Prediction

Price Column Before Outliers Taken Out
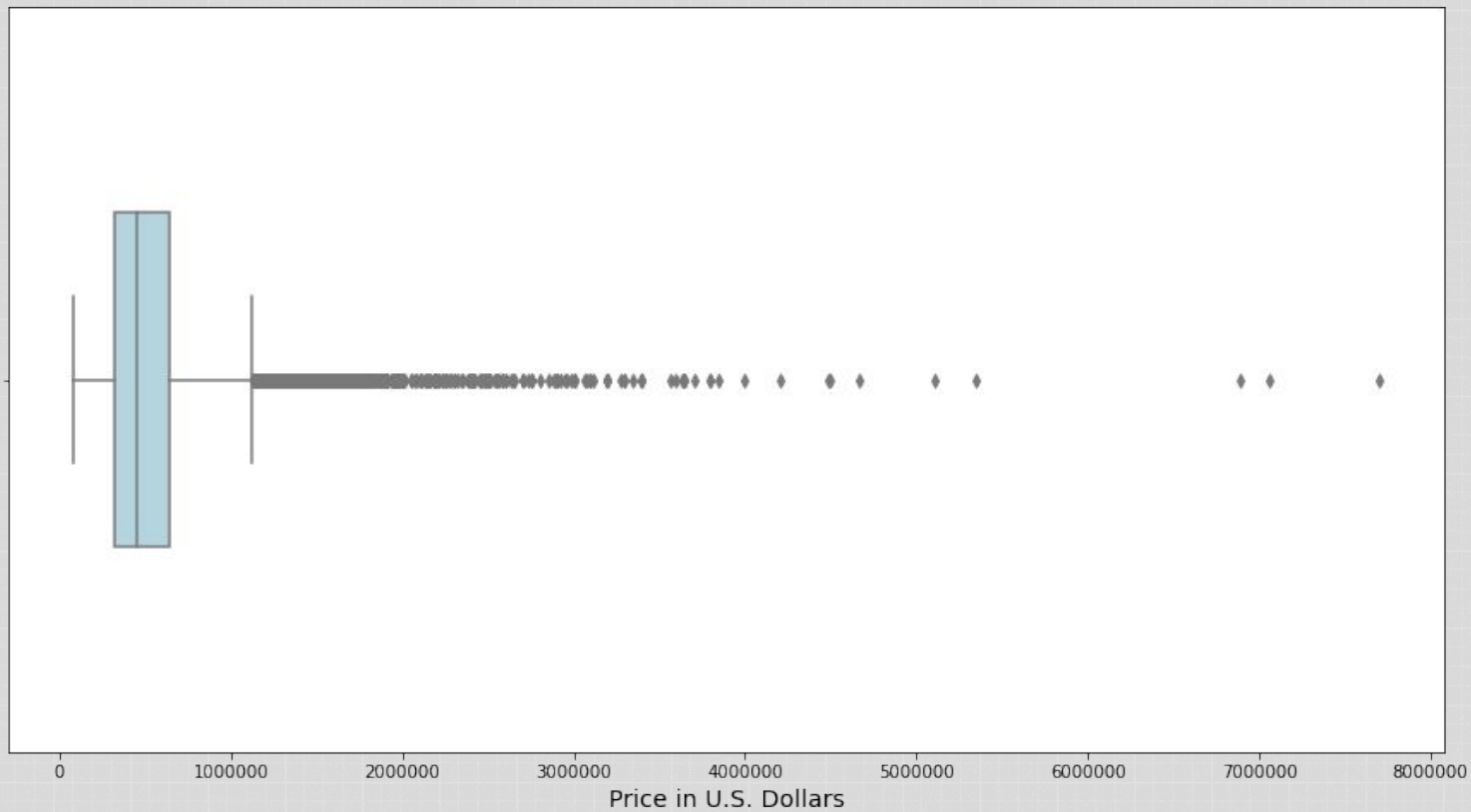
After Outliers Taken Out

Prices

Better Way to Clean Outliers

```python
from scipy.stats import zscore
from scipy.stats import stats
z_scores = stats.zscore(df2.price)
abs_z_scores = np.abs(z_scores)
filtered_entries = (abs_z_scores < 3)
new_df2_price = df2.price[filtered_entries]
new_df2_price
```

```python
q_low = df_2["price"].quantile(0.01)
q_hi  = df_2["price"].quantile(0.99)

df_filtered = df_2[(df_2["price"] < q_hi) & (df_2["price"] > q_low)]
```
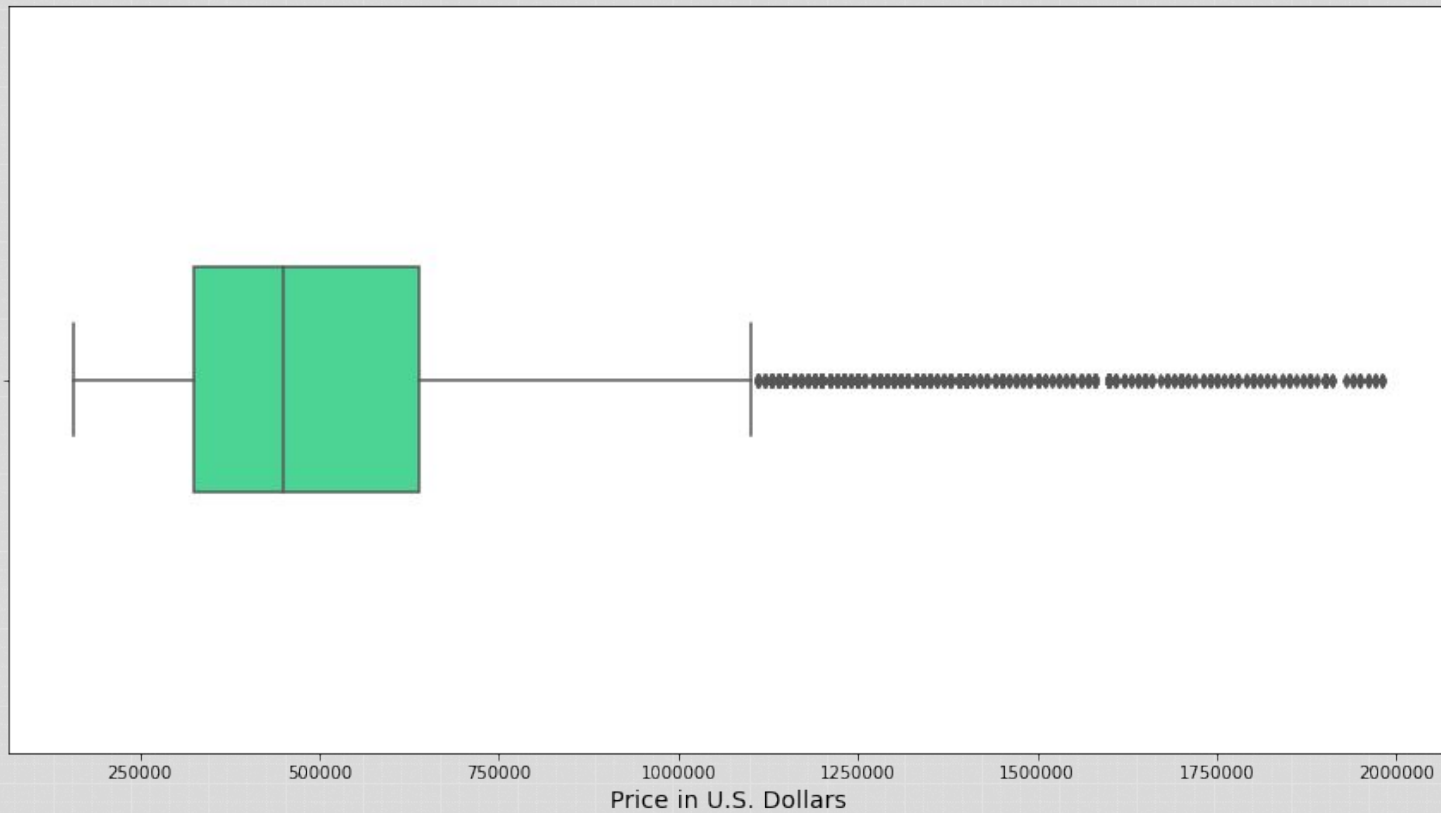
Before Outliers Taken Out

Price in U.S. Dollars

After Outliers Taken Out

Price in U.S. Dollars

Joe's   vs.   Katie's

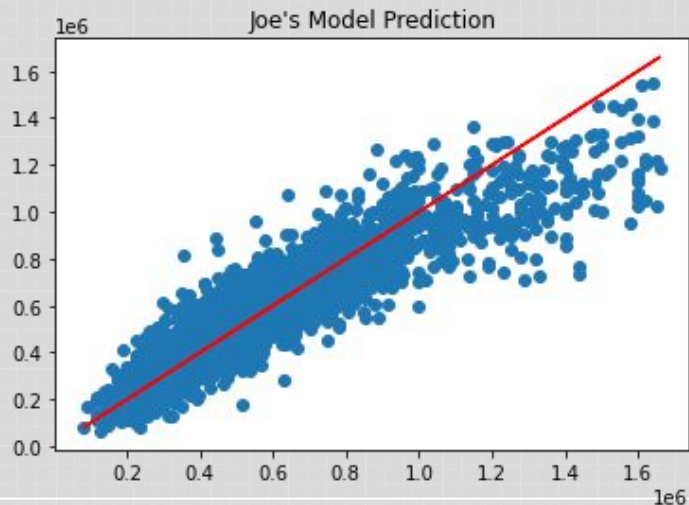Training R^2: 0.8477

Testing R^2: 0.8473

Training RMSE: 101155.68

Testing RMSE: 105763.01

Training R^2: 0.8493

Testing R^2: 0.8557

Training RMSE: 109665.08

Testing RMSE: 109613.52

Joe's Model Prediction

Preprocessing

- Dropped year renovated column

- Did not improve the model once

Preprocessing

- Bathrooms are categorical, not continuous.

- Included every column except id, but tried encoding bathrooms along with the other variables from earlier.
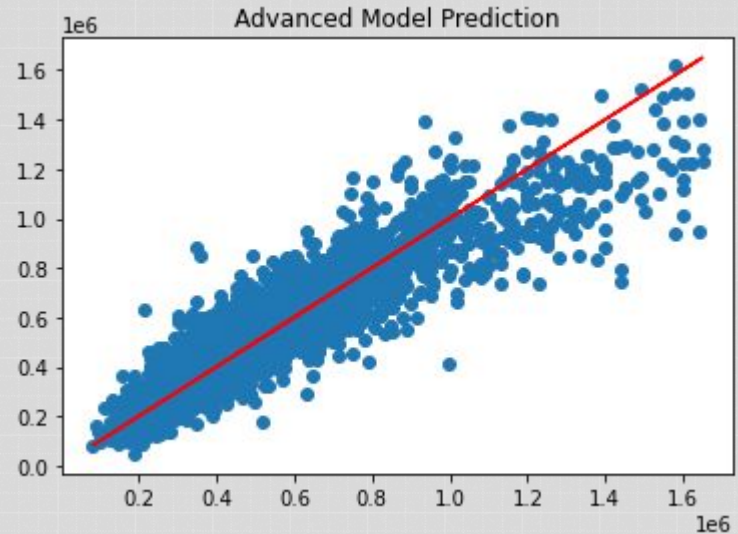
- Improved the model

Advanced Model

Training R^2: 0.8503

Testing R^2: 0.8467

Training RMSE: 101415.6435

Testing RMSE: 102664.7123

Advanced Model Prediction

If We Had More Time...

Create algorithms to account for each house owner's individual experience

Prospective house buyers could input values for various features and would receive an estimate of how much a house with those features would cost.

THANK YOU!

Katie Skoda
Email: kjskoda@bsc.edu
GitHub: @kjskoda

Joe Swing
Email: joeswing88@gmail.com
GitHub: @jswing450

Colette Crowder
Email: crcrowde@bsc.edu
GitHub: @crcrowde