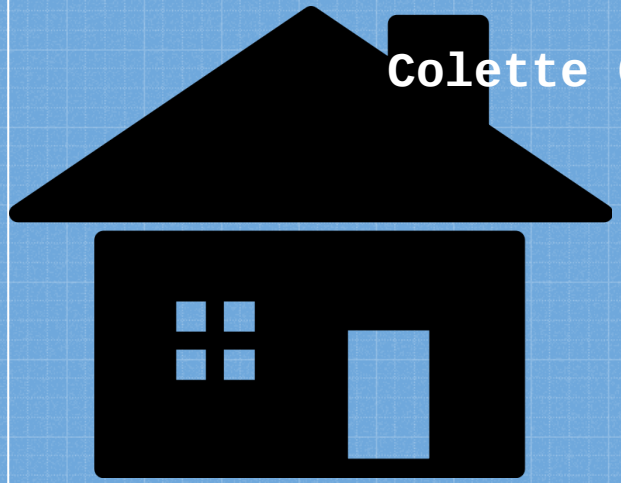
King County Real Estate Predictive Modeling

Colette Crowder, Joe Swing, Katie Skoda

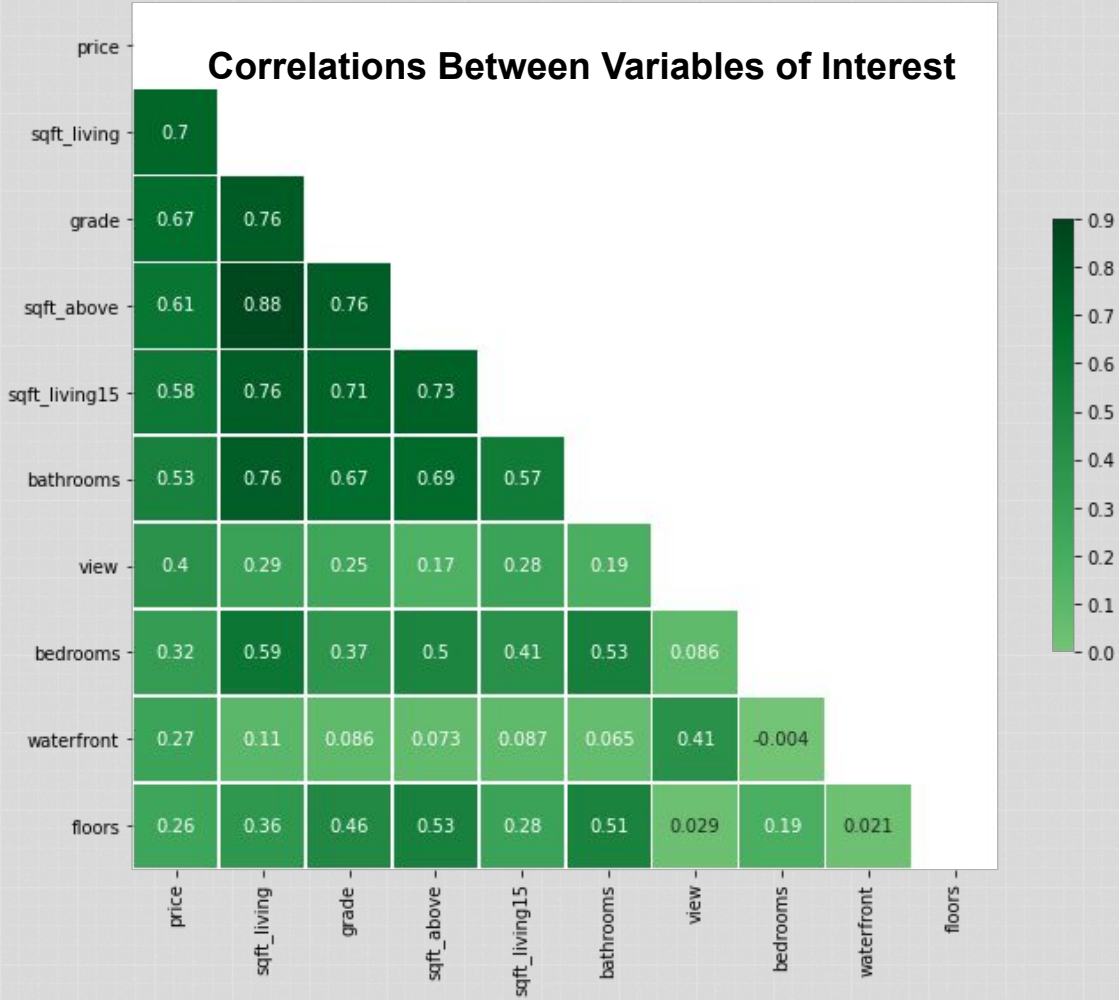


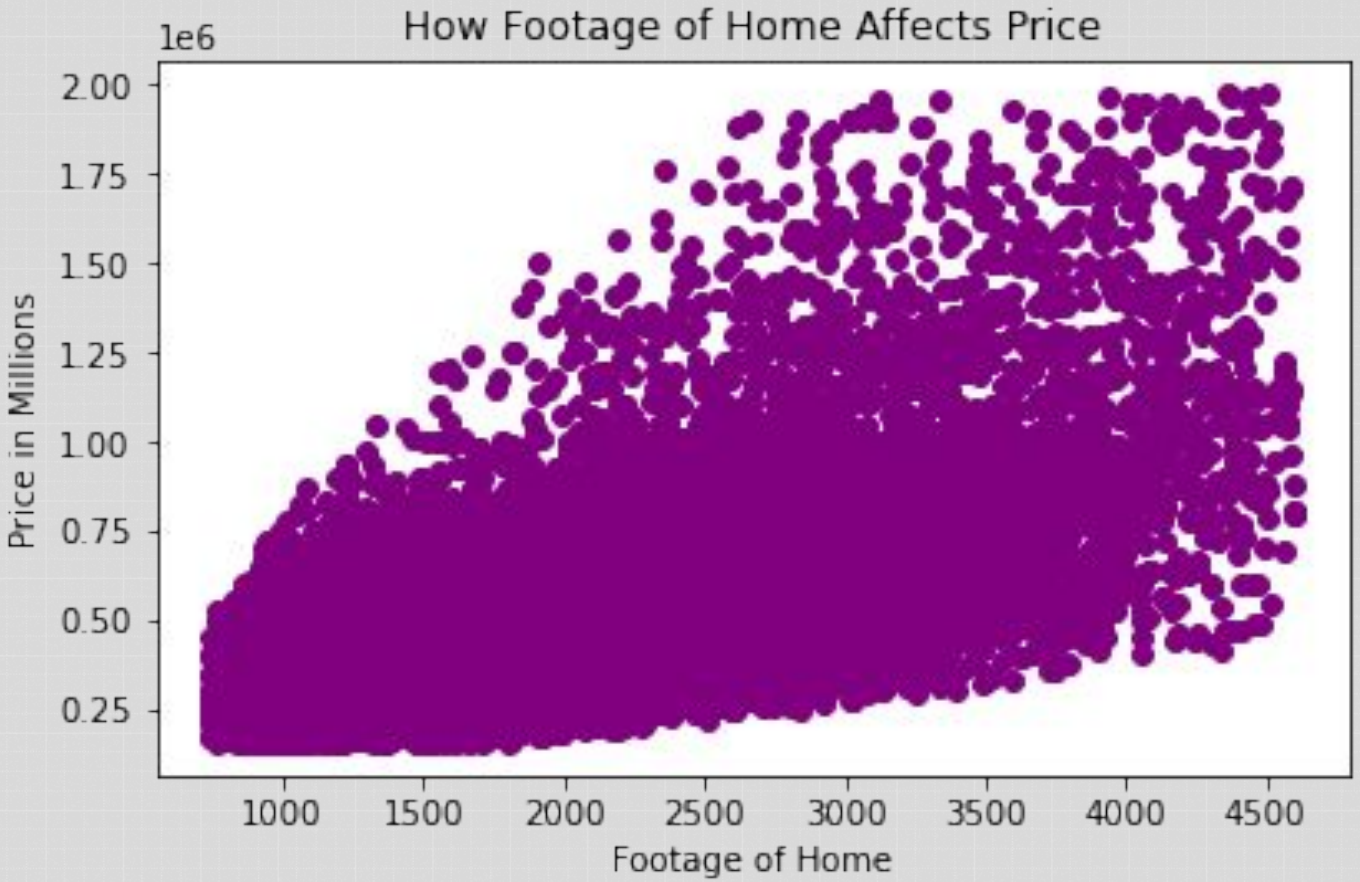
Cleaning

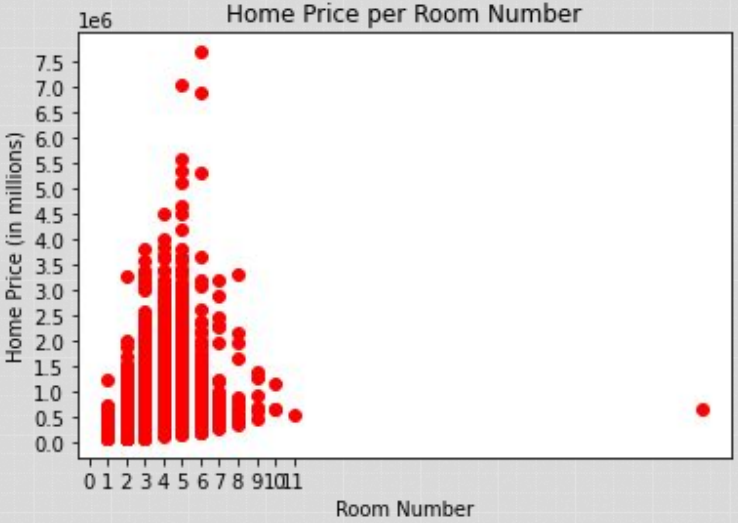
- I replaced all ‘?’ with NaN values

- 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

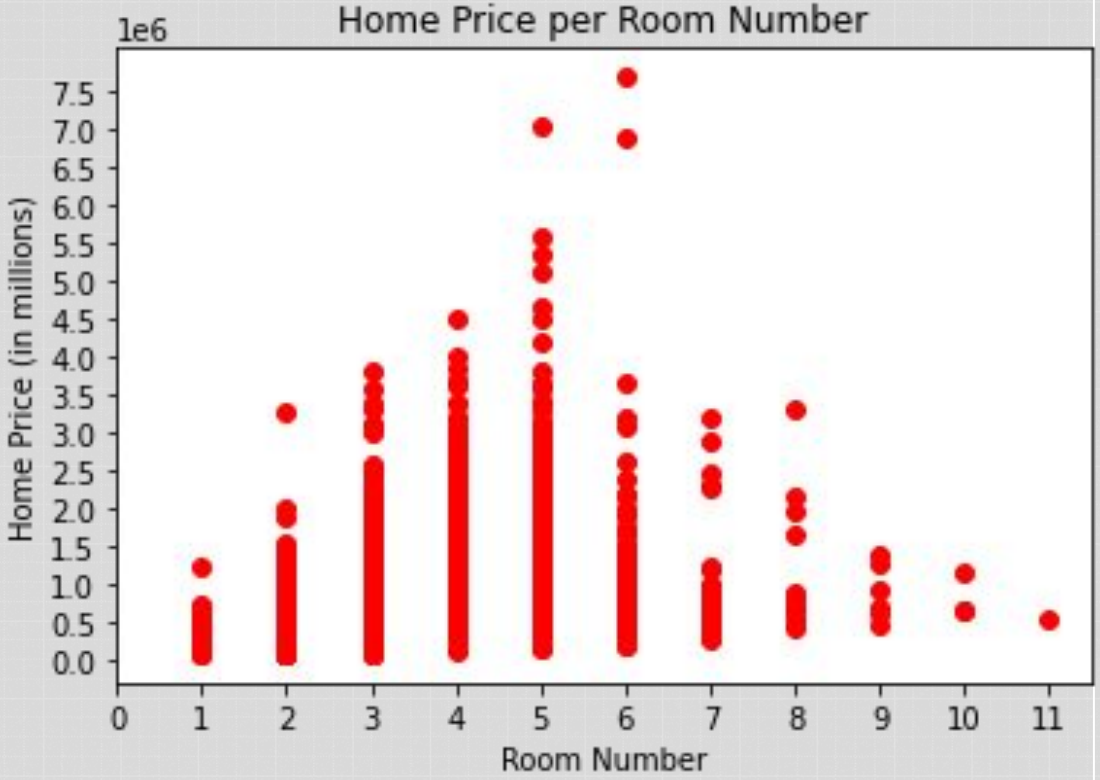
- Removed the entry that had 33 bedrooms
- Turned date from a string to a int using just the year
- Encoded view and grade for our baseline model


Correlations Between Variables of Interest
price
sqft_living
grade
sqft_above
sqft_living15
bathrooms
view
bedrooms
waterfront
floors
0.7
0.67 0.76
0.61 0.88 0.76
0.58 0.76 0.71 0.73
0.53 0.76 0.67 0.69 0.57
0.4 0.29 0.25 0.17 0.28 0.19
0.32 0.59 0.37 0.5 0.41 0.53 0.086
0.27 0.11 0.086 0.073 0.087 0.065 0.41 -0.004
0.26 0.36 0.46 0.53 0.28 0.51 0.029 0.19 0.021
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0


How Footage of Home Affects Price
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
Price in Millions
1000
1500
2000
2500
3000
3500
4000
4500
Footage of Home



Dropped the outlier of 33 bedrooms for a better visualization and more precise data



Baseline Model

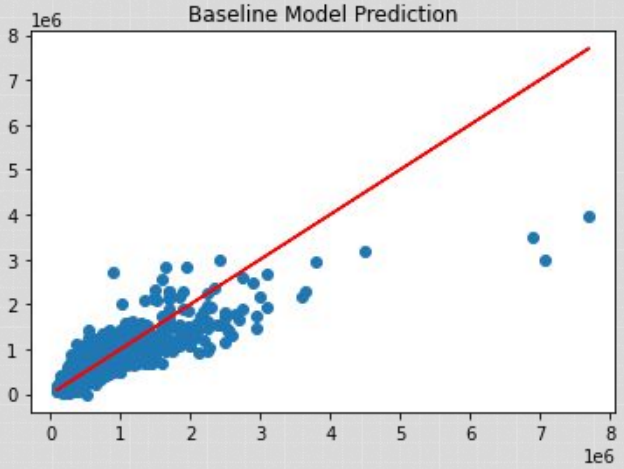
- For the baseline model, we included every variable except for id.
- The scores we got for our testing and training data were:

Training R^2: 0.7339

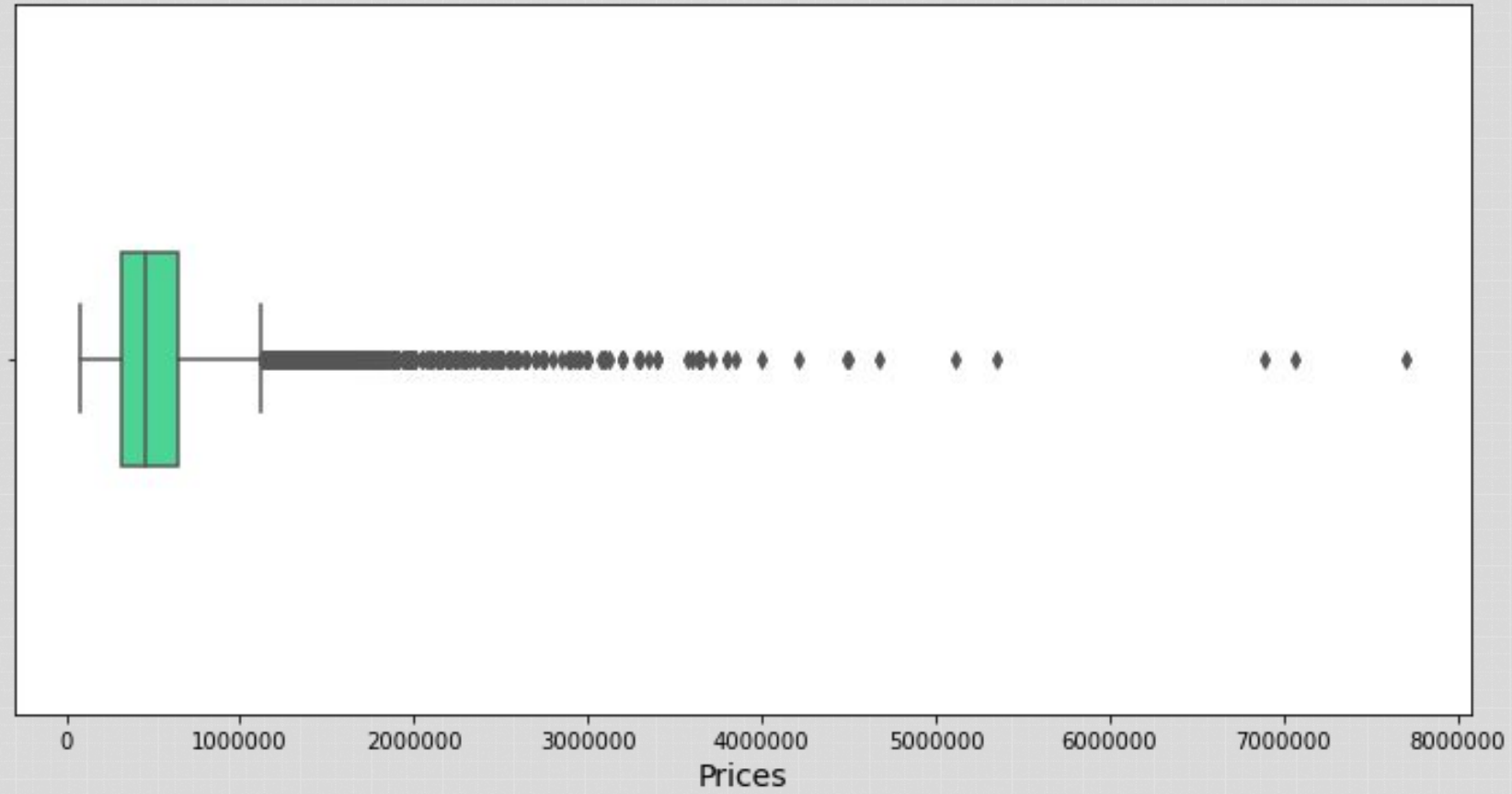
Testing R^2: 0.7184

Training RMSE: 185625.8124

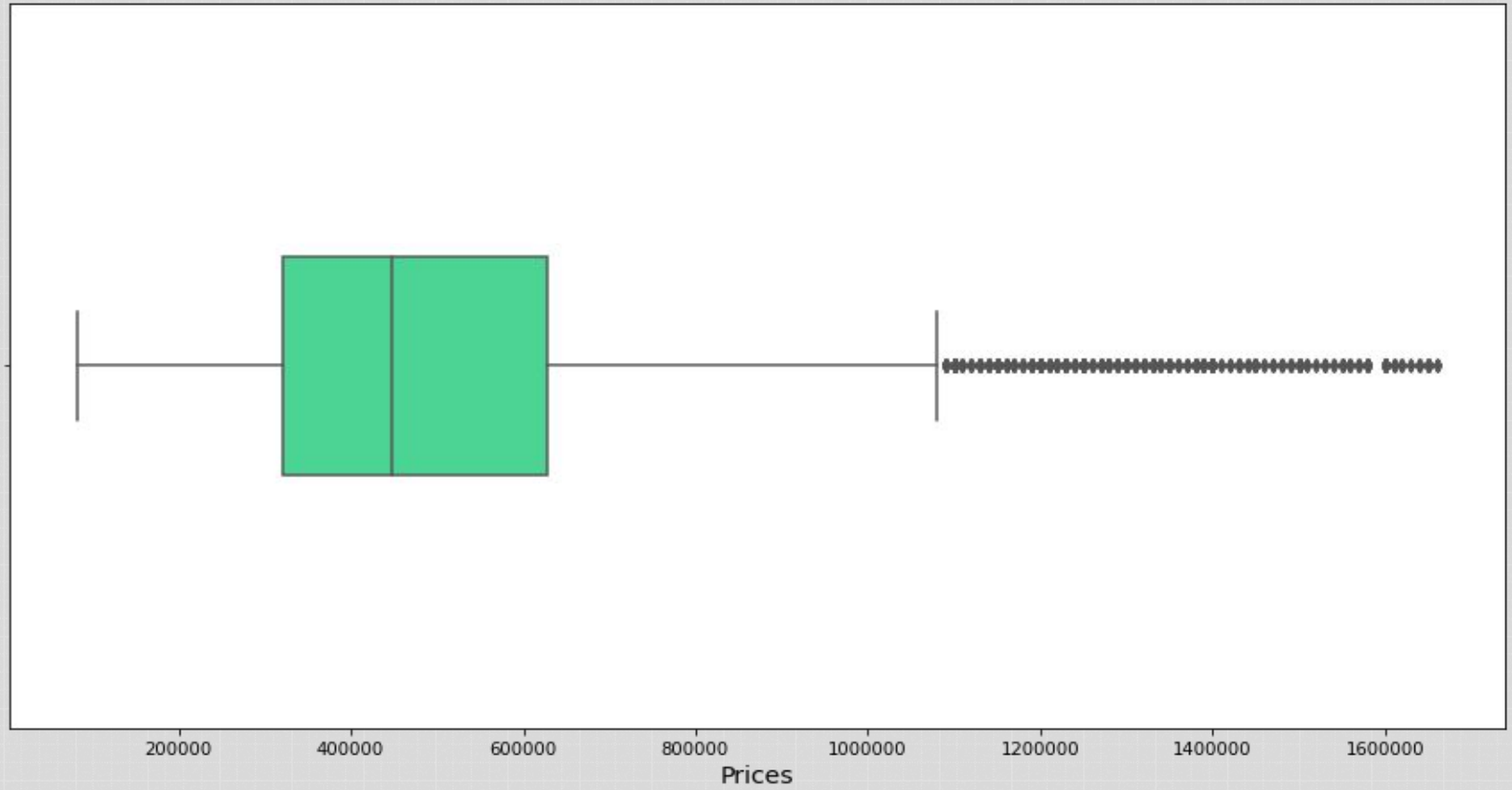
Testing RMSE: 217398.4034



Price Column Before Outliers Taken Out



After Outliers Taken Out



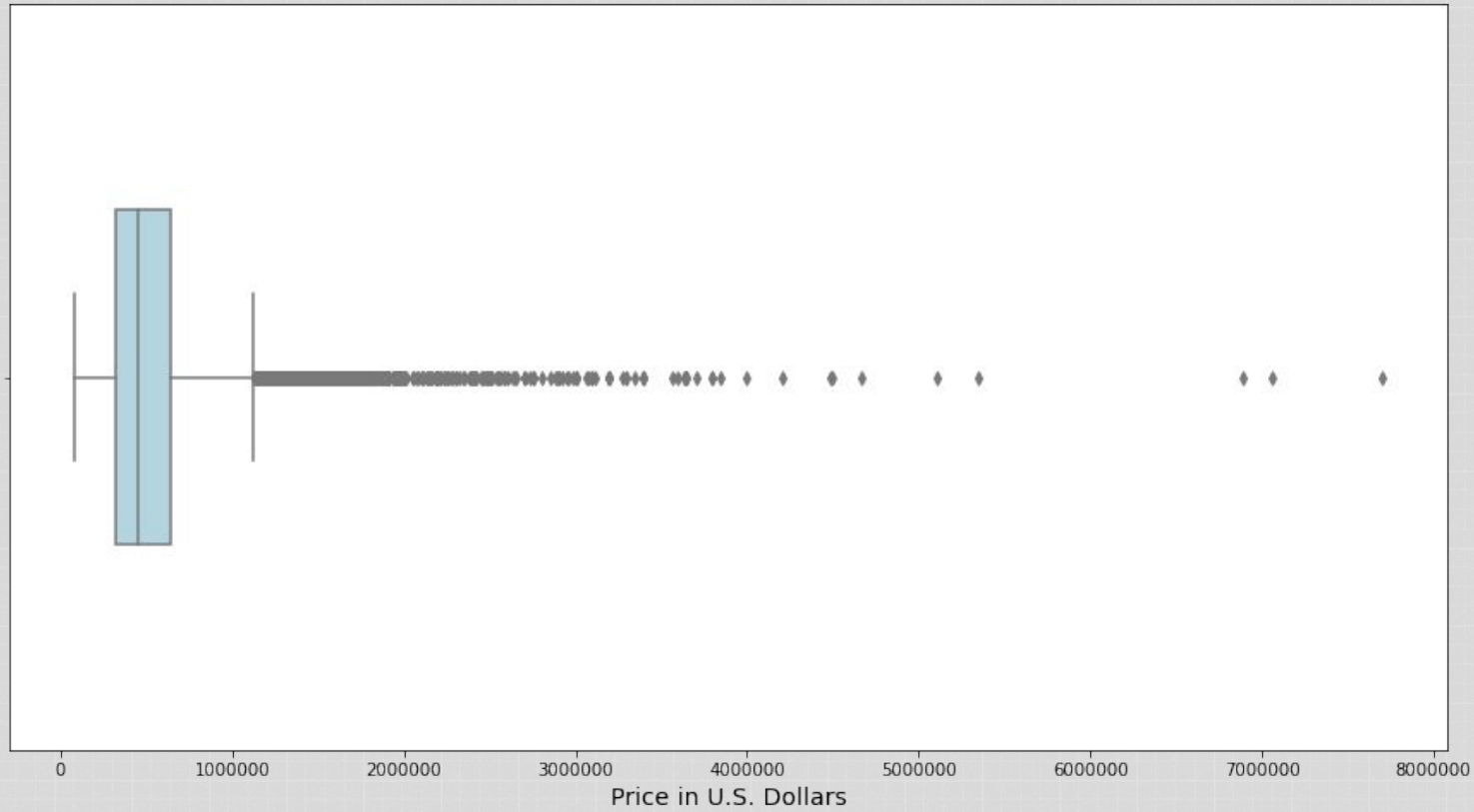
Better Way to Clean Outliers

```
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
```

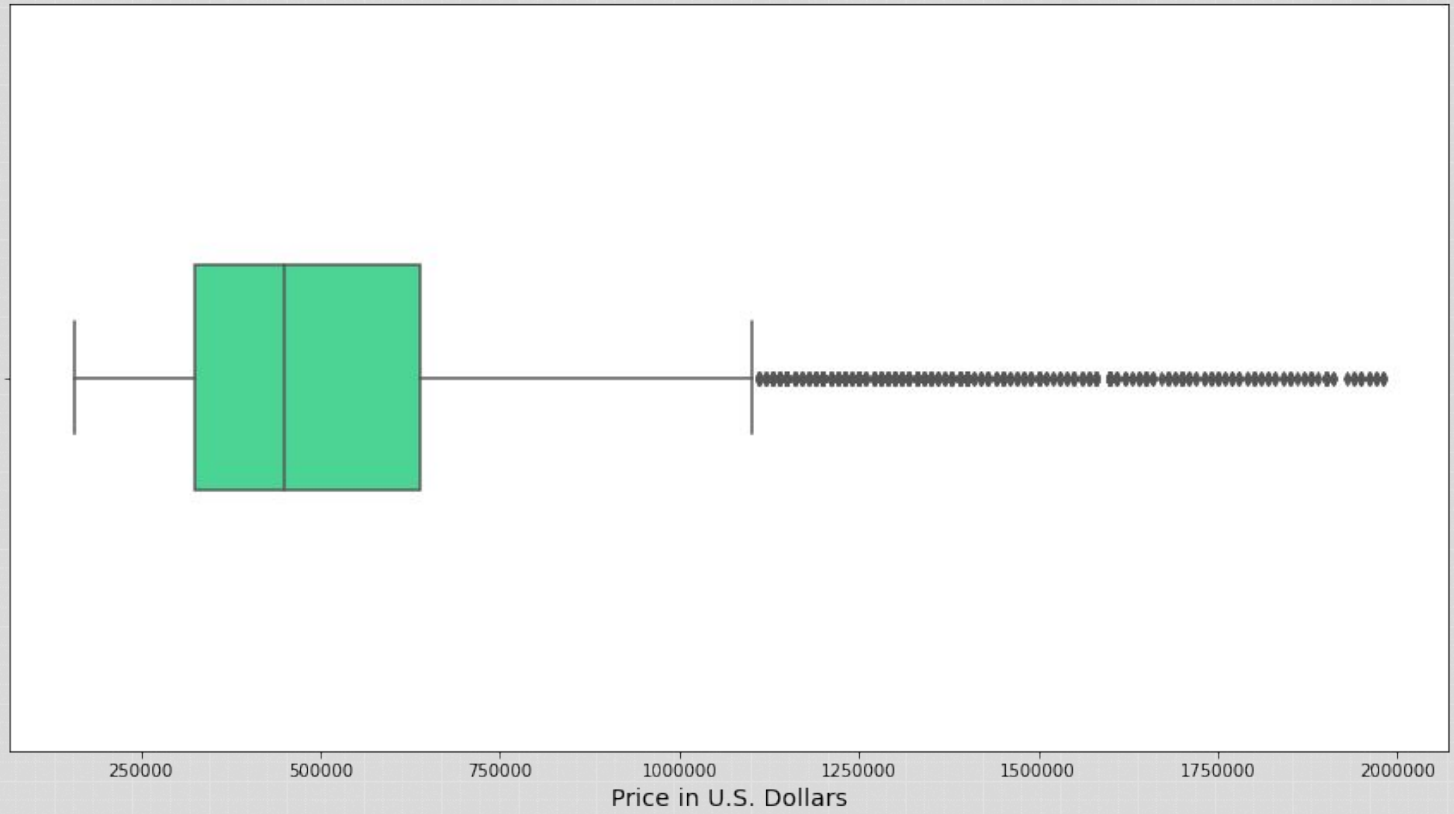
```
q_low = df_2["price"].quantile(0.01)
q_hi  = df_2["price"].quantile(0.99)

df_filtered = df_2[(df_2["price"] < q_hi) & (df_2["price"] > q_low)]
```

Before Outliers Taken Out



After Outliers Taken Out



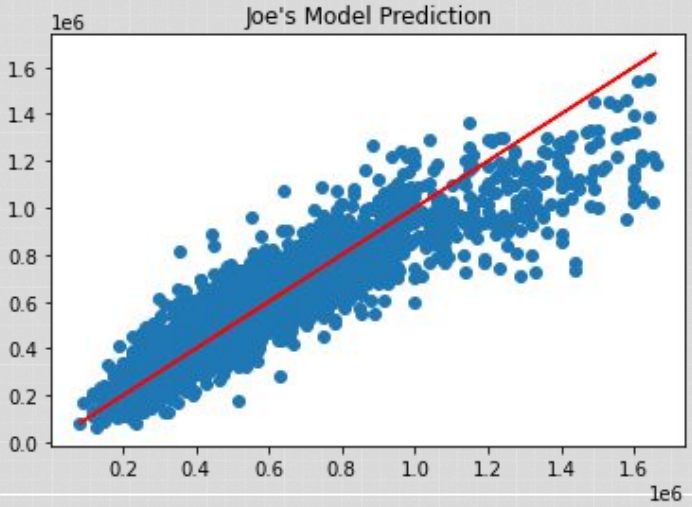
Joe's vs. Katie's

Training R^2: 0.8477

Testing R^2: 0.8473

Training RMSE: 101155.68

Testing RMSE: 105763.01

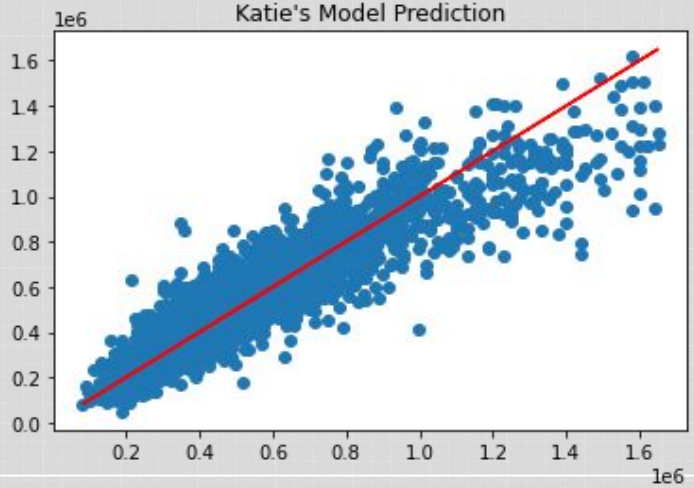


Training R^2: 0.8493

Testing R^2: 0.8557

Training RMSE: 109665.08

Testing RMSE: 109613.52



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 zip code, view, and grade from earlier.
- Improved the model

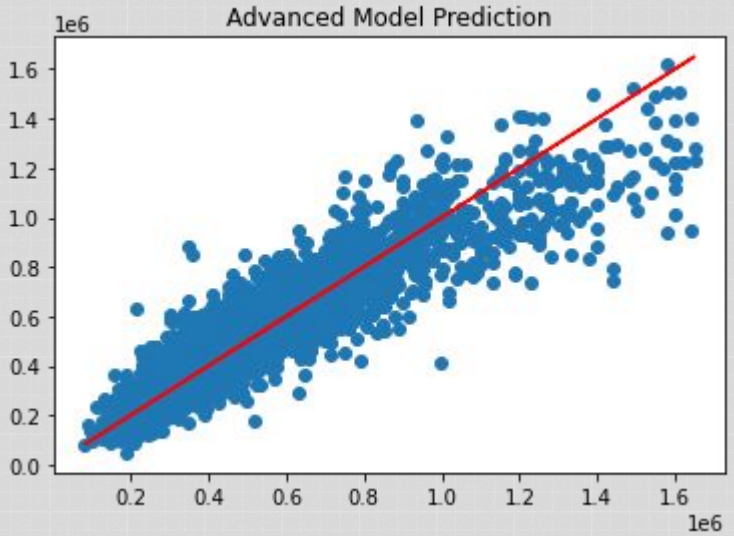
Advanced Model

Training R^2: 0.8503

Testing R^2: 0.8467

Training RMSE: 101415.6435

Testing RMSE: 102664.7123



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