Predicting Home Prices

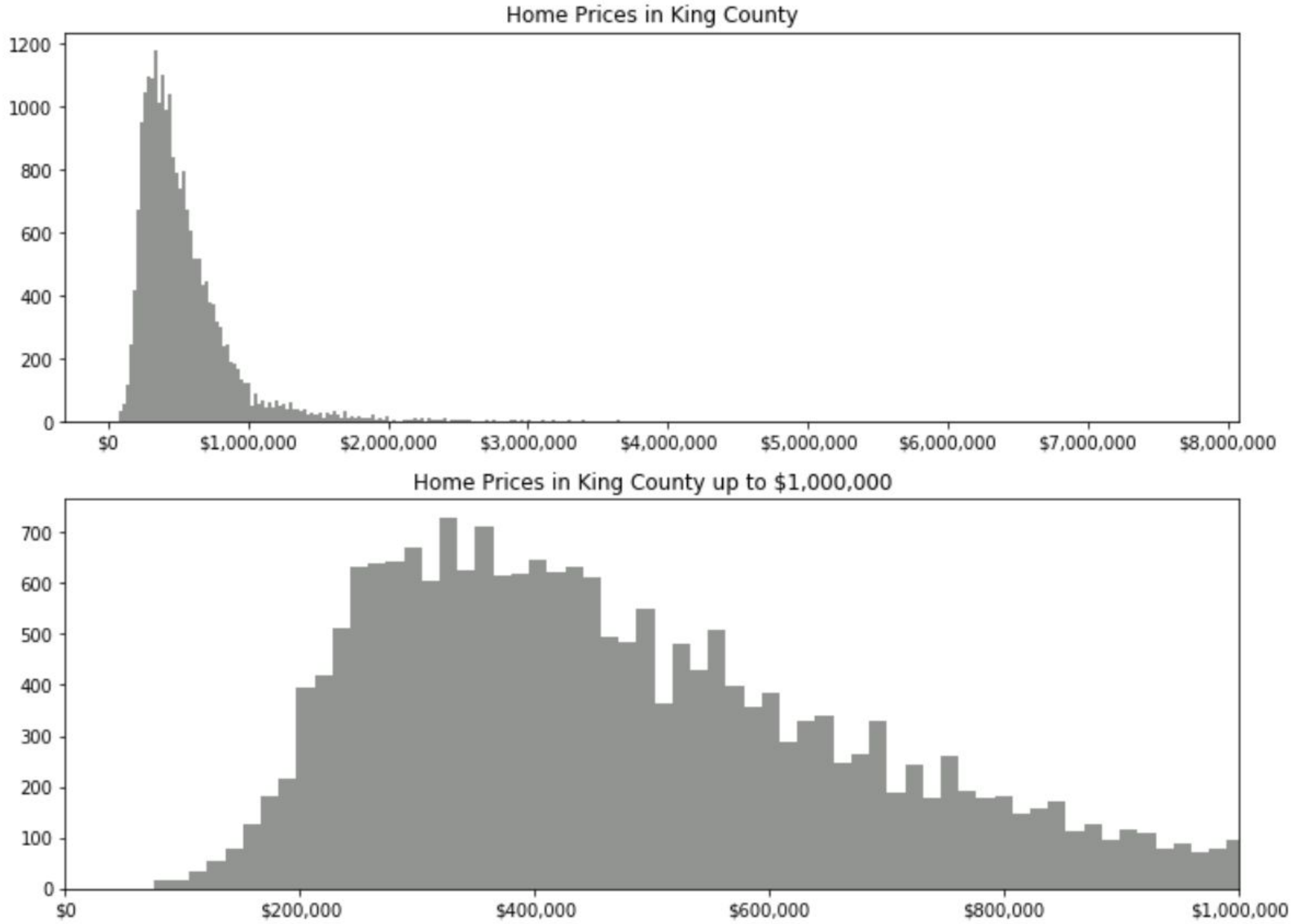
King County, Washington

- What are the greatest influencers in the pricing of homes?
- Explore data visually
- Create linear regression model to predict home prices

The Dataset

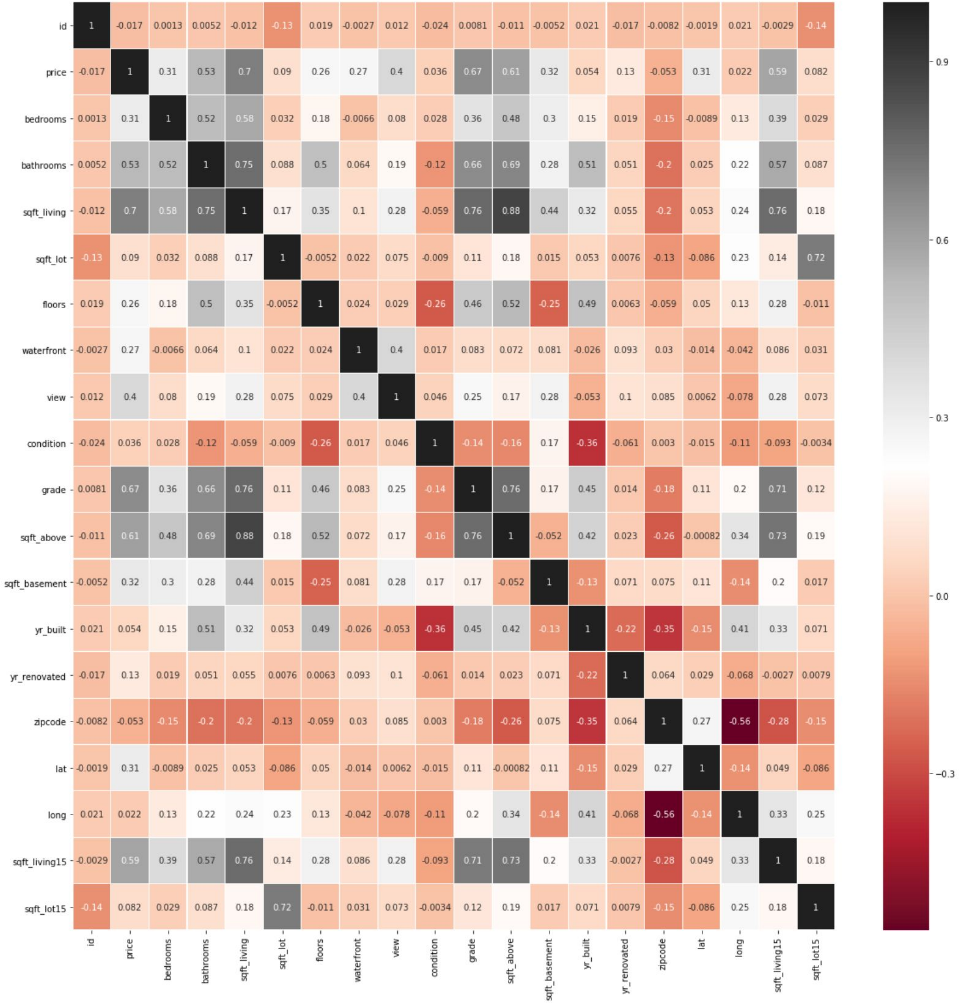
The dataset was made available on Kaggle.com by user harlfoxem. The data is clean and ready for analysis.

https://www.kaggle.com/harlfoxem/housesalesprediction/data



Home Prices in King County

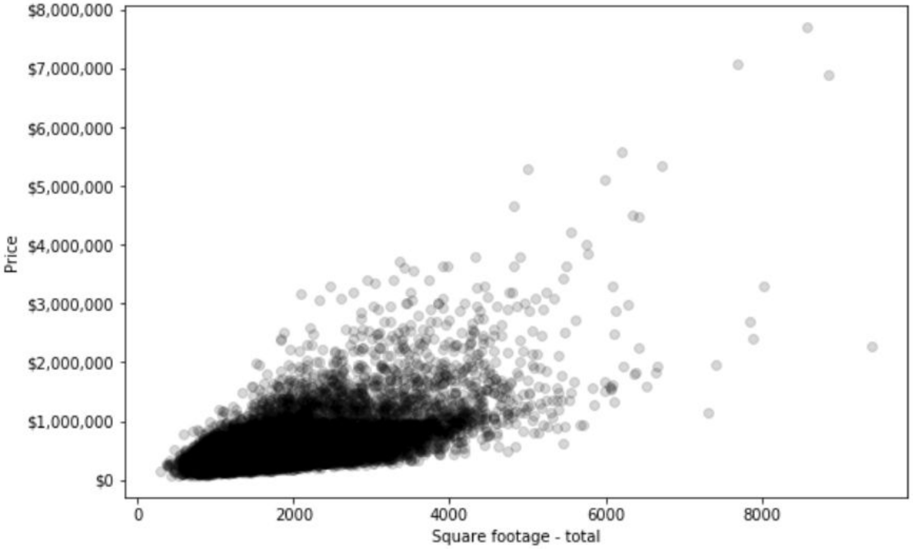
Home Prices in King County up to $1,000,000

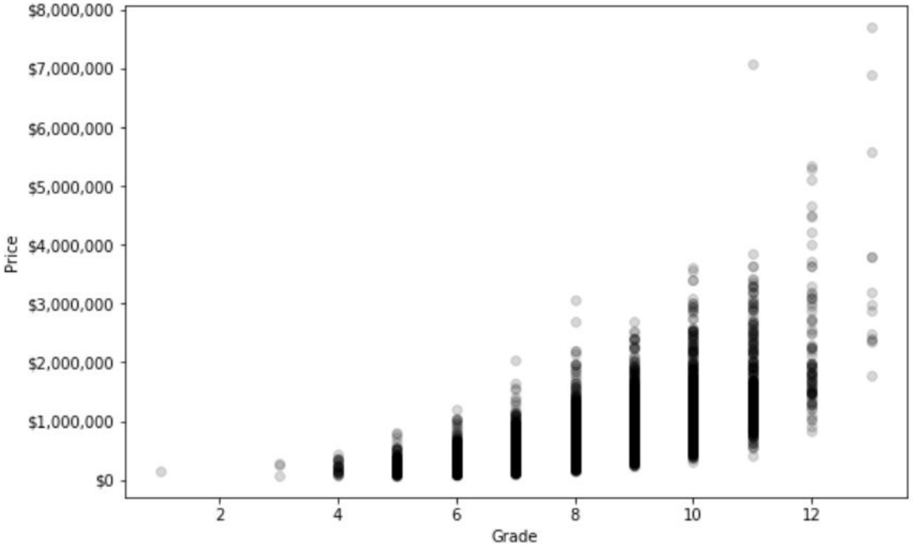


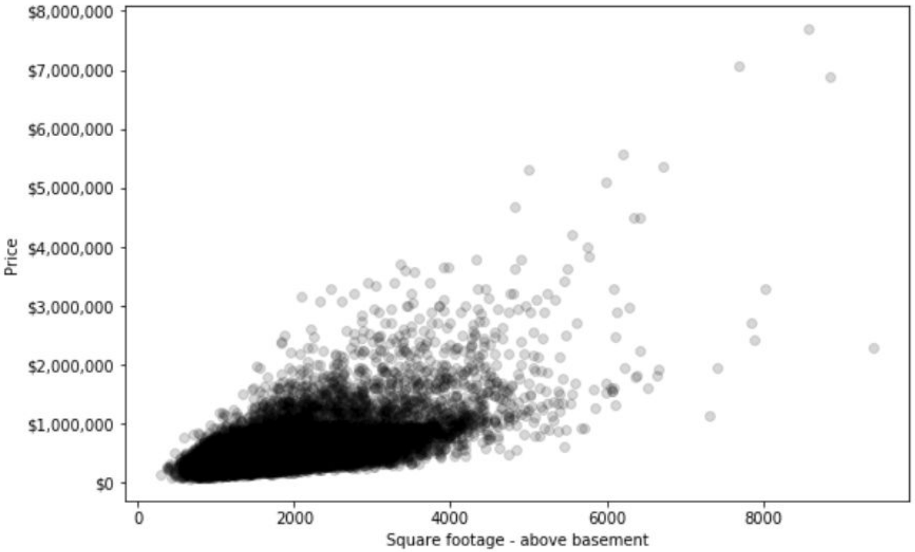
	id	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built	yr_renovated	zipcode	lat	long	sqft_living15	sqft_lot15
id	1	-0.017	0.0013	0.0052	-0.012	-0.13	0.019	-0.0027	0.012	-0.024	0.0081	-0.011	-0.0052	0.021	-0.017	-0.0082	-0.0019	0.021	-0.0029	-0.14
price	-0.017	1	0.31	0.53	0.7	0.09	0.26	0.27	0.4	0.036	0.67	0.61	0.32	0.054	0.13	-0.053	0.31	0.022	0.59	0.082
bedrooms	0.0013	0.31	1	0.52	0.58	0.032	0.18	-0.0066	0.08	0.028	0.36	0.48	0.3	0.15	0.019	-0.15	-0.0089	0.13	0.39	0.029
bathrooms	0.0052	0.53	0.52	1	0.75	0.088	0.5	0.064	0.19	-0.12	0.66	0.69	0.28	0.51	0.051	-0.2	0.025	0.22	0.57	0.087
sqft_living	-0.012	0.7	0.58	0.75	1	0.17	0.35	0.1	0.28	-0.059	0.76	0.88	0.44	0.32	0.055	-0.2	0.053	0.24	0.76	0.18
sqft_lot	-0.13	0.09	0.032	0.088	0.17	1	-0.0052	0.022	0.075	-0.009	0.11	0.18	0.015	0.053	0.0076	-0.13	-0.086	0.23	0.14	0.72
floors	0.019	0.26	0.18	0.5	0.35	-0.0052	1	0.024	0.029	-0.26	0.46	0.52	-0.25	0.49	0.0063	-0.059	0.05	0.13	0.28	-0.011
waterfront	-0.0027	0.27	-0.0066	0.064	0.1	0.022	0.024	1	0.4	0.017	0.083	0.072	0.081	-0.026	0.093	0.03	-0.014	-0.042	0.086	0.031
view	0.012	0.4	0.08	0.19	0.28	0.075	0.029	0.4	1	0.046	0.25	0.17	0.28	-0.053	0.1	0.085	0.0062	-0.078	0.28	0.073
condition	-0.024	0.036	0.028	-0.12	-0.059	-0.009	-0.26	0.017	0.046	1	-0.14	-0.16	0.17	-0.36	-0.061	0.003	-0.015	-0.11	-0.093	-0.0034
grade	0.0081	0.67	0.36	0.66	0.76	0.11	0.46	0.083	0.25	-0.14	1	0.76	0.17	0.45	0.014	-0.18	0.11	0.2	0.71	0.12
sqft_above	-0.011	0.61	0.48	0.69	0.88	0.18	0.52	0.072	0.17	-0.16	0.76	1	-0.052	0.42	0.023	-0.26	-0.00082	0.34	0.73	0.19
sqft_basement	-0.0052	0.32	0.3	0.28	0.44	0.015	-0.25	0.081	0.28	0.17	0.17	-0.052	1	-0.13	0.071	0.075	0.11	-0.14	0.2	0.017
yr_built	0.021	0.054	0.15	0.51	0.32	0.053	0.49	-0.026	-0.053	-0.36	0.45	0.42	-0.13	1	-0.22	-0.35	-0.15	0.41	0.33	0.071
yr_renovated	-0.017	0.13	0.019	0.051	0.055	0.0076	0.0063	0.093	0.1	-0.061	0.014	0.023	0.071	-0.22	1	0.064	0.029	-0.068	-0.0027	0.0079
zipcode	-0.0082	-0.053	-0.15	-0.2	-0.2	-0.13	-0.059	0.03	0.085	0.003	-0.18	-0.26	0.075	-0.35	0.064	1	0.27	-0.56	-0.28	-0.15
lat	-0.0019	0.31	-0.0089	0.025	0.053	-0.086	0.05	-0.014	0.0062	-0.015	0.11	-0.00082	0.11	-0.15	0.029	0.27	1	-0.14	0.049	-0.086
long	0.021	0.022	0.13	0.22	0.24	0.23	0.13	-0.042	-0.078	-0.11	0.2	0.34	-0.14	0.41	-0.068	-0.56	-0.14	1	0.33	0.25
sqft_living15	-0.0029	0.59	0.39	0.57	0.76	0.14	0.28	0.086	0.28	-0.093	0.71	0.73	0.2	0.33	-0.0027	-0.28	0.049	0.33	1	0.18
sqft_lot15	-0.14	0.082	0.029	0.087	0.18	0.72	-0.011	0.031	0.073	-0.0034	0.12	0.19	0.017	0.071	0.0079	-0.15	-0.086	0.25	0.18	1

Correlates to price

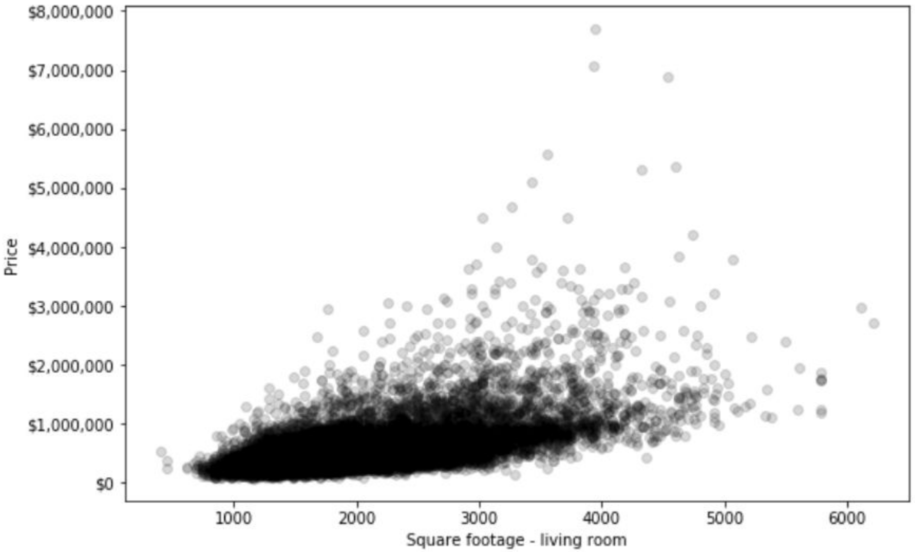
sqft_living	0.702035
grade	0.667434
sqft_above	0.605567
sqft_living15	0.585379
bathrooms	0.525138
view	0.397293
sqft_basement	0.323816
bedrooms	0.308350
lat	0.307003
waterfront	0.266369


$8,000,000
$7,000,000
$6,000,000
$5,000,000
$4,000,000
$3,000,000
$2,000,000
$1,000,000
$0
Price
0
2000
4000
6000
8000
Square footage - total

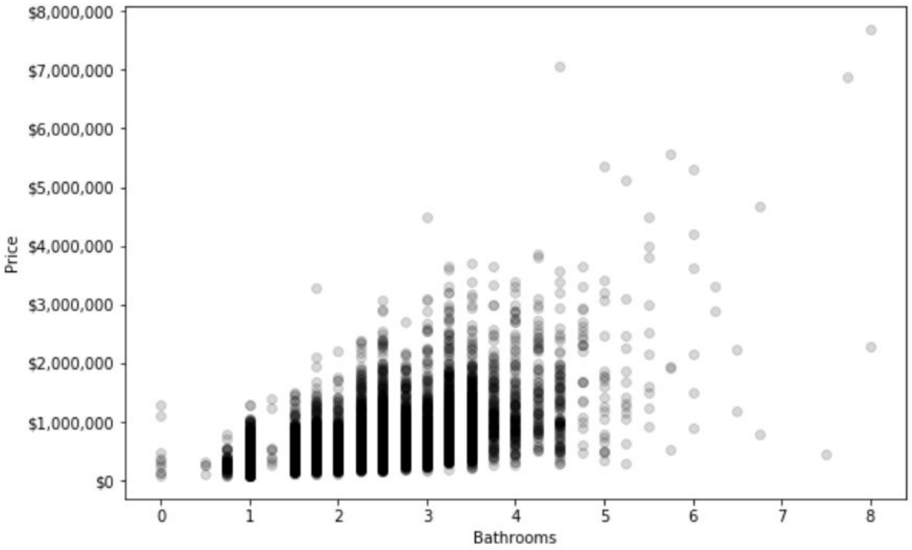

$8,000,000
$7,000,000
$6,000,000
$5,000,000
$4,000,000
$3,000,000
$2,000,000
$1,000,000
$0
Price
2
4
6
8
10
12
Grade

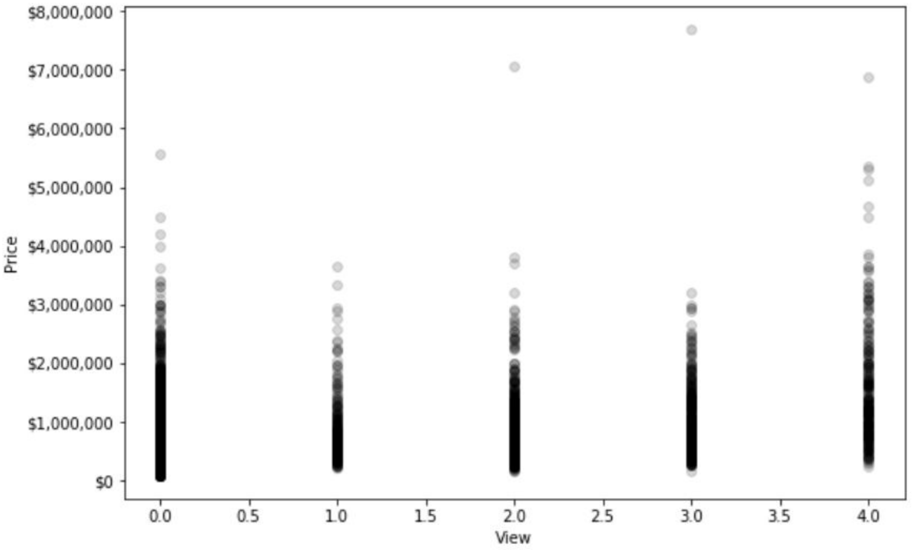


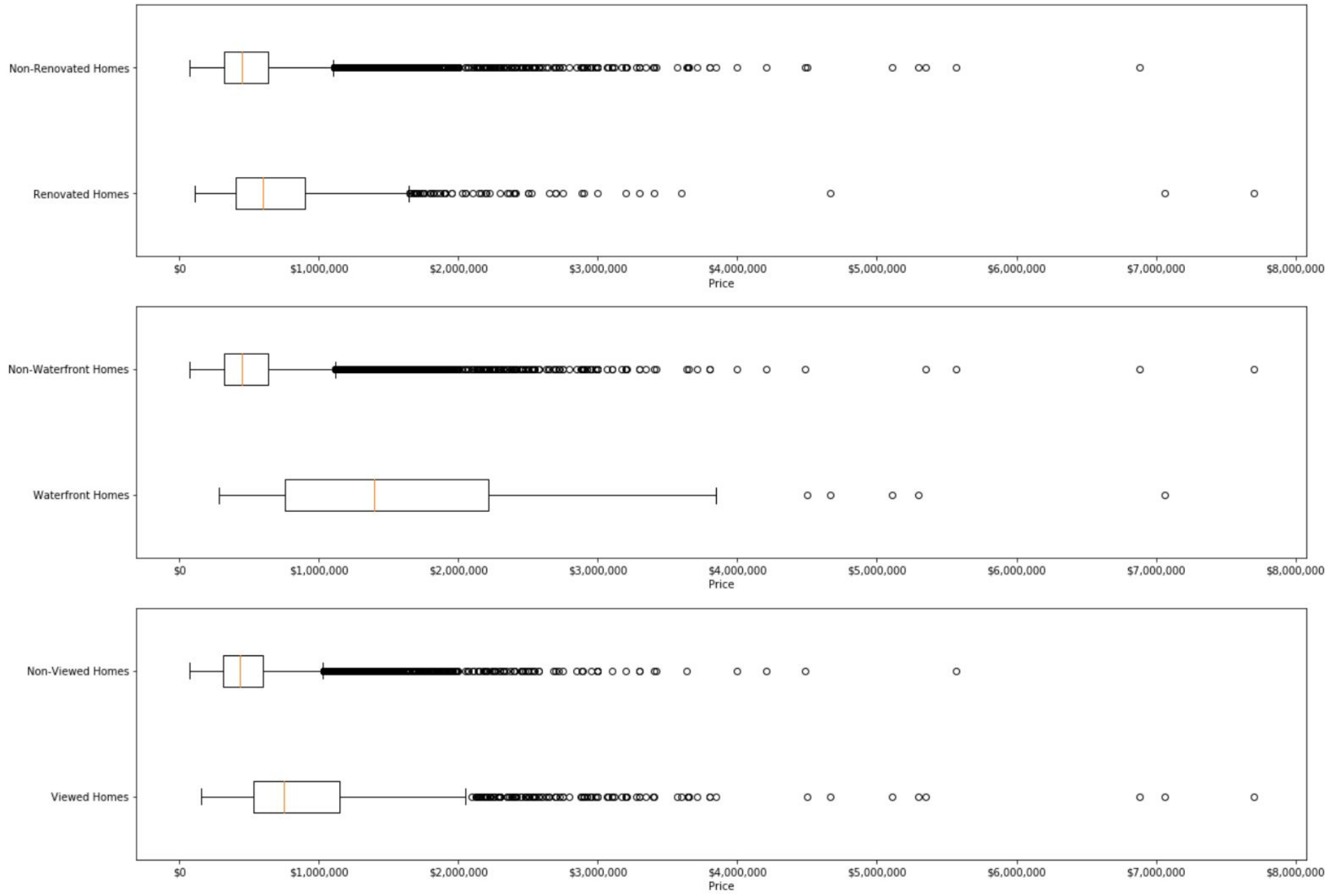
$8,000,000
$7,000,000
$6,000,000
$5,000,000
$4,000,000
$3,000,000
$2,000,000
$1,000,000
$0
Price
0
2000
4000
6000
8000
Square footage - above basement



$8,000,000
$7,000,000
$6,000,000
$5,000,000
$4,000,000
$3,000,000
$2,000,000
$1,000,000
$0
Price
1000
2000
3000
4000
5000
6000
Square footage - living room


$8,000,000
$7,000,000
$6,000,000
$5,000,000
$4,000,000
$3,000,000
$2,000,000
$1,000,000
$0
Price
0
1
2
3
4
5
6
7
8
Bathrooms

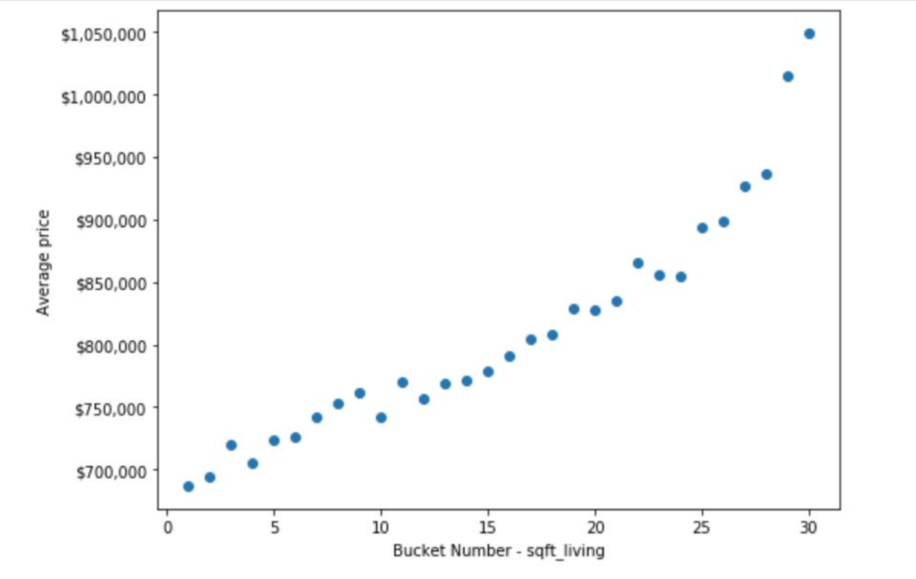

$8,000,000
$7,000,000
$6,000,000
$5,000,000
$4,000,000
$3,000,000
$2,000,000
$1,000,000
$0
Price
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
View


Non-Renovated Homes
Renovated Homes
$0
$1,000,000
$2,000,000
$3,000,000
$4,000,000
$5,000,000
$6,000,000
$7,000,000
$8,000,000
Price
Non-Waterfront Homes
Waterfront Homes
$0
$1,000,000
$2,000,000
$3,000,000
$4,000,000
$5,000,000
$6,000,000
$7,000,000
$8,000,000
Price
Non-Viewed Homes
Viewed Homes
$0
$1,000,000
$2,000,000
$3,000,000
$4,000,000
$5,000,000
$6,000,000
$7,000,000
$8,000,000
Price

Feature Encoding

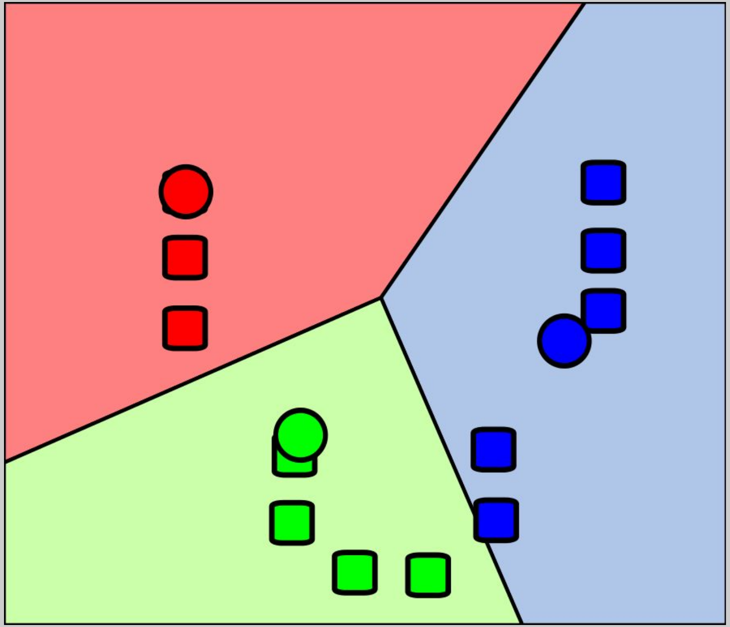
To eliminate some of the variability of a feature, we can group each observation into intervals of similarly priced homes. After each observation is in it's bucket, we calculate the average of each bucket, and then assign that average value to a new column in the dataframe.

For example, when I encode the feature 'sqft_living', each home will be grouped into 30 equally populated buckets corresponding to their square footage. I'll calculate the average of each bucket, and it will be assigned to a new feature, which can be used to predict the price.

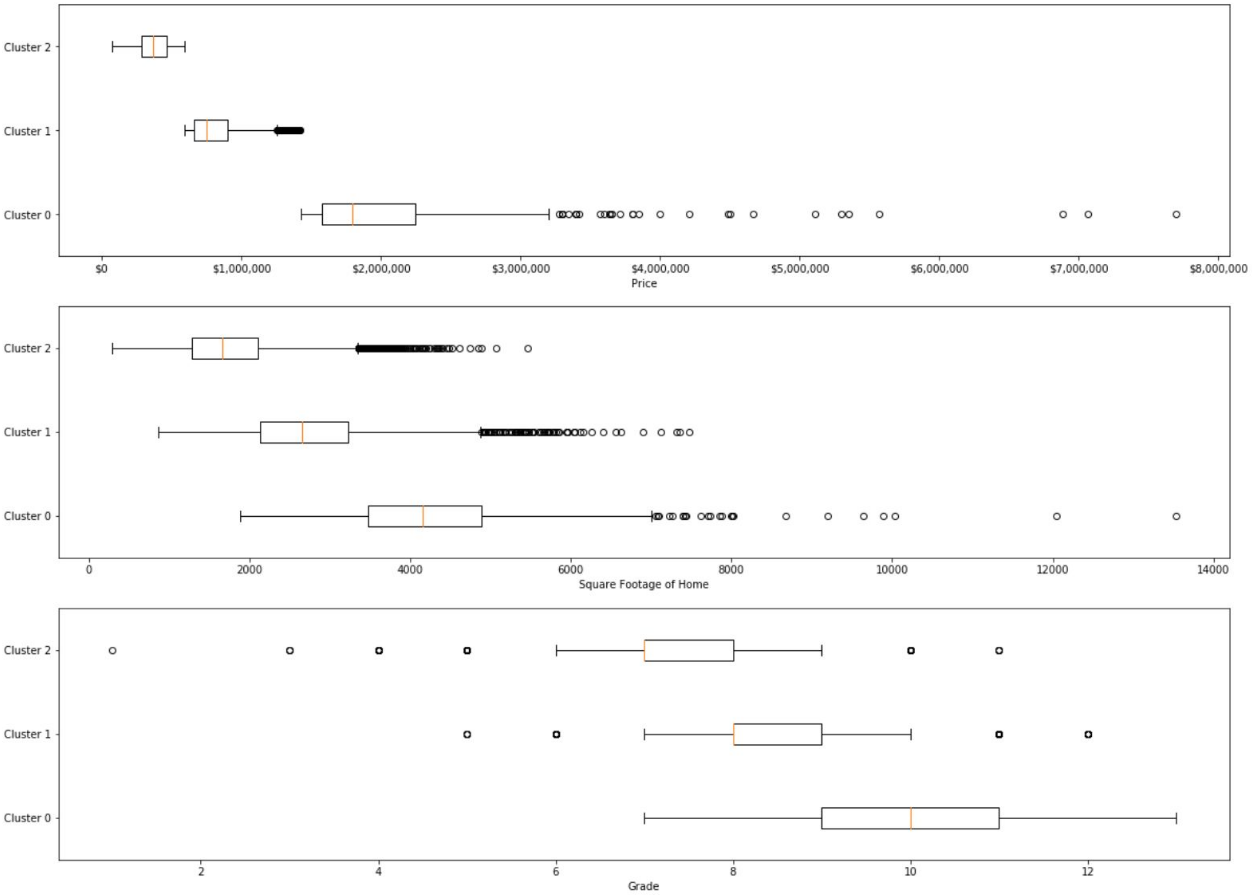


Clustering

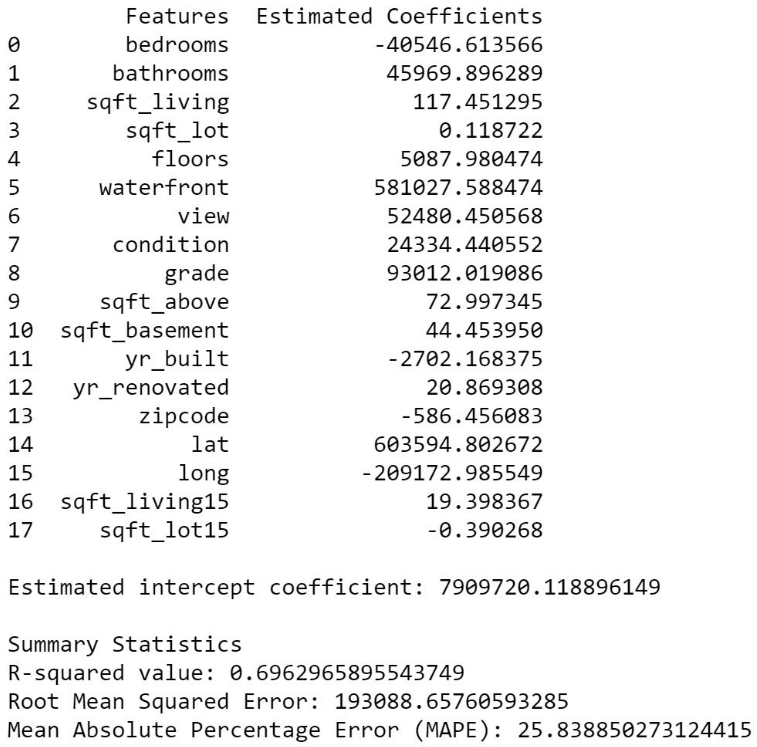
Kmeans clustering is an unsupervised learning technique used to identify similar groupings in a dataset. We can use clustering to find groups of homes that have similar characteristics, and create models of these smaller subgroups to obtain a more accurate overall model.



Results of Clustering



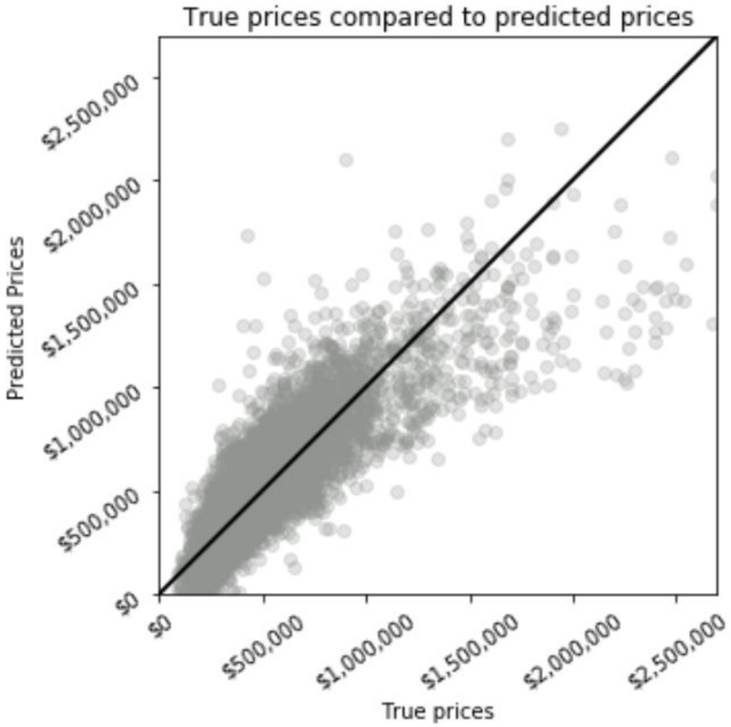
Model using all features



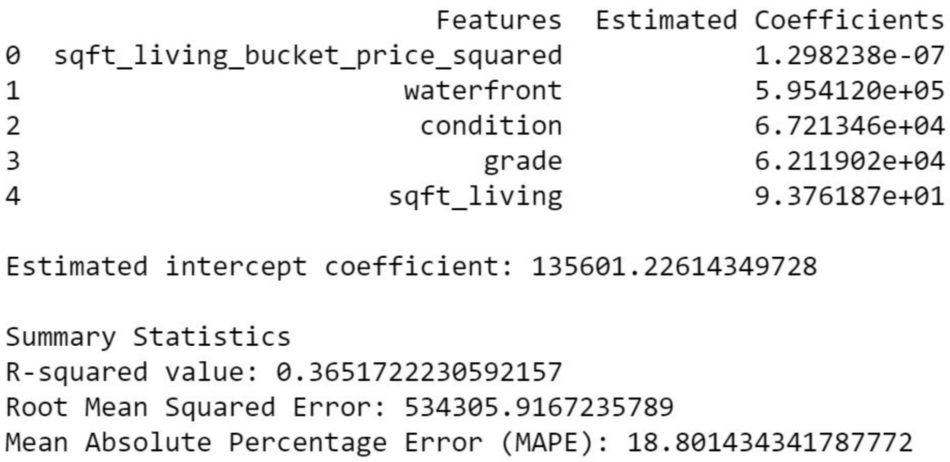
```
          Features  Estimated Coefficients
0         bedrooms           -40546.613566
1        bathrooms            45969.896289
2      sqft_living              117.451295
3         sqft_lot                0.118722
4           floors             5087.980474
5       waterfront           581027.588474
6             view            52480.450568
7        condition            24334.440552
8            grade            93012.019086
9       sqft_above               72.997345
10   sqft_basement               44.453950
11        yr_built            -2702.168375
12    yr_renovated               20.869308
13         zipcode             -586.456083
14             lat           603594.802672
15            long          -209172.985549
16   sqft_living15               19.398367
17      sqft_lot15               -0.390268

Estimated intercept coefficient: 7909720.118896149

Summary Statistics
R-squared value: 0.6962965895543749
Root Mean Squared Error: 193088.65760593285
Mean Absolute Percentage Error (MAPE): 25.838850273124415
```



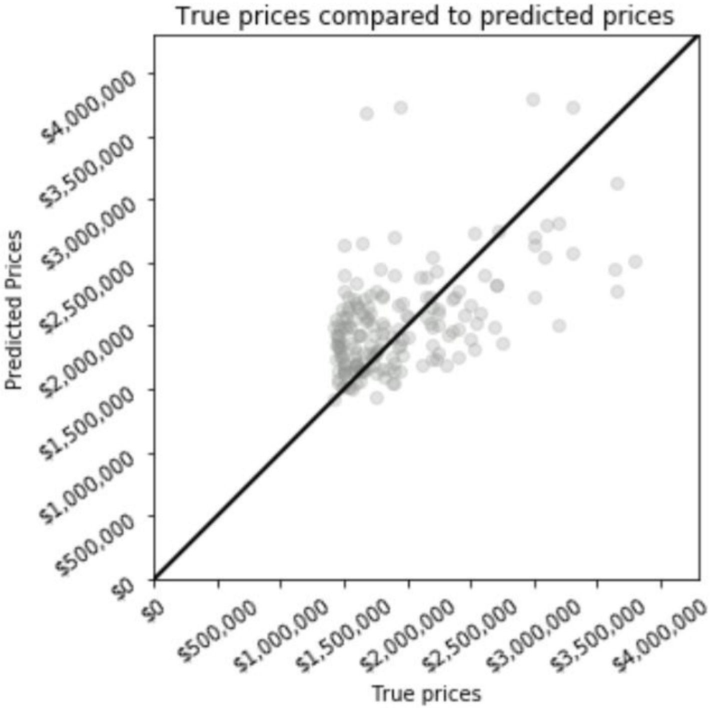
Cluster 0 Model



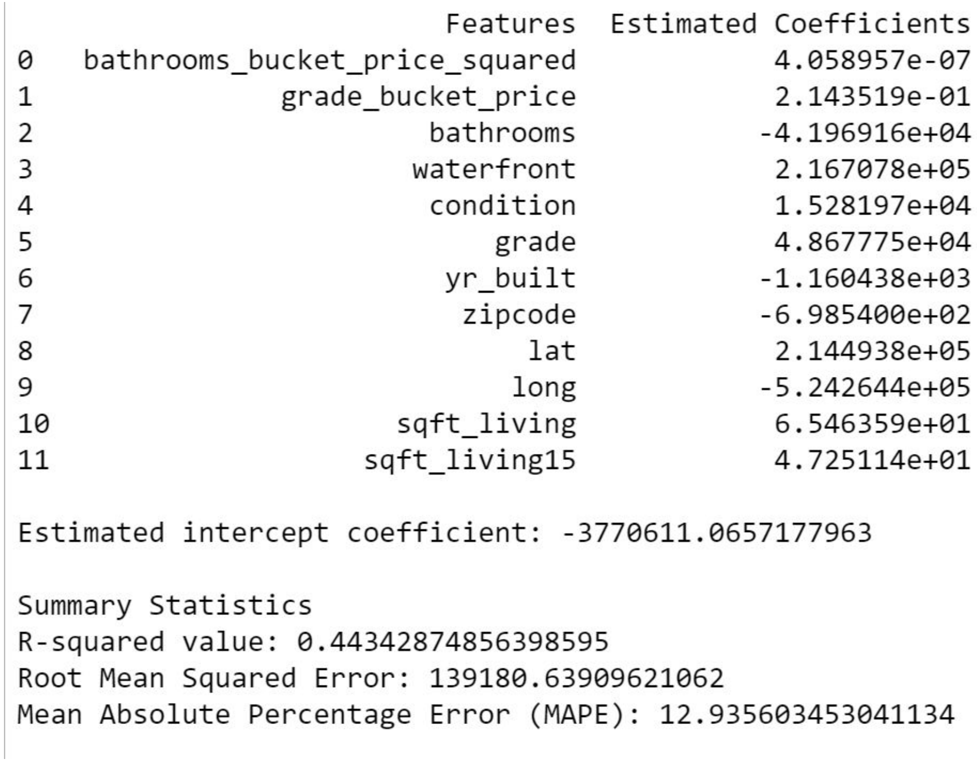
```
                          Features  Estimated Coefficients
0  sqft_living_bucket_price_squared            1.298238e-07
1                        waterfront            5.954120e+05
2                         condition            6.721346e+04
3                             grade            6.211902e+04
4                       sqft_living            9.376187e+01

Estimated intercept coefficient: 135601.22614349728

Summary Statistics
R-squared value: 0.3651722230592157
Root Mean Squared Error: 534305.9167235789
Mean Absolute Percentage Error (MAPE): 18.801434341787772
```



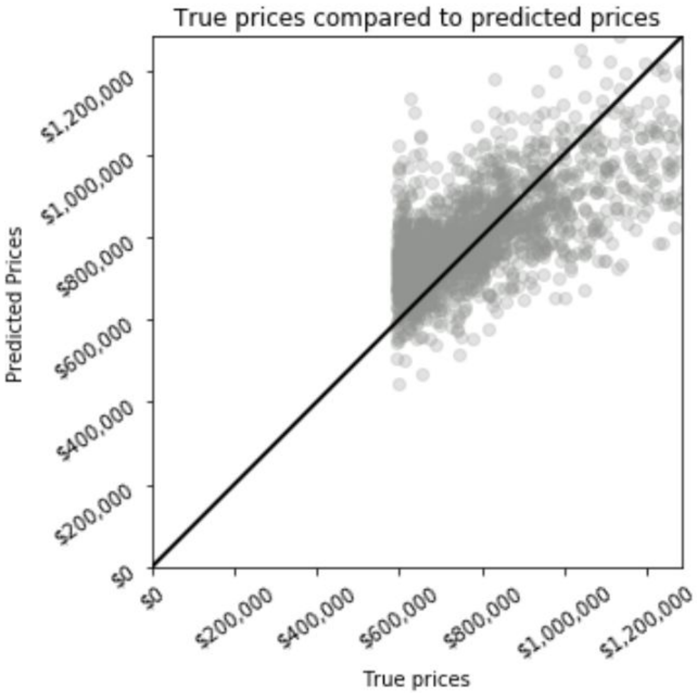
Cluster 1 Model



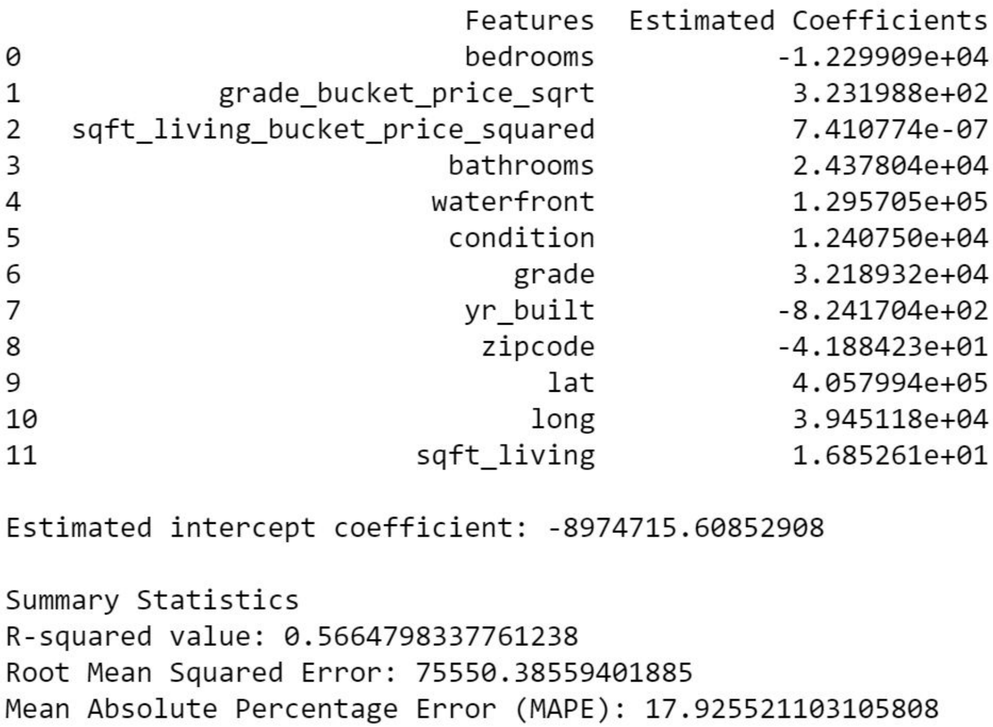
```
                    Features  Estimated Coefficients
0   bathrooms_bucket_price_squared          4.058957e-07
1             grade_bucket_price          2.143519e-01
2                      bathrooms         -4.196916e+04
3                     waterfront          2.167078e+05
4                      condition          1.528197e+04
5                          grade          4.867775e+04
6                       yr_built         -1.160438e+03
7                        zipcode         -6.985400e+02
8                            lat          2.144938e+05
9                           long         -5.242644e+05
10                   sqft_living          6.546359e+01
11                 sqft_living15          4.725114e+01

Estimated intercept coefficient: -3770611.0657177963

Summary Statistics
R-squared value: 0.44342874856398595
Root Mean Squared Error: 139180.63909621062
Mean Absolute Percentage Error (MAPE): 12.935603453041134
```



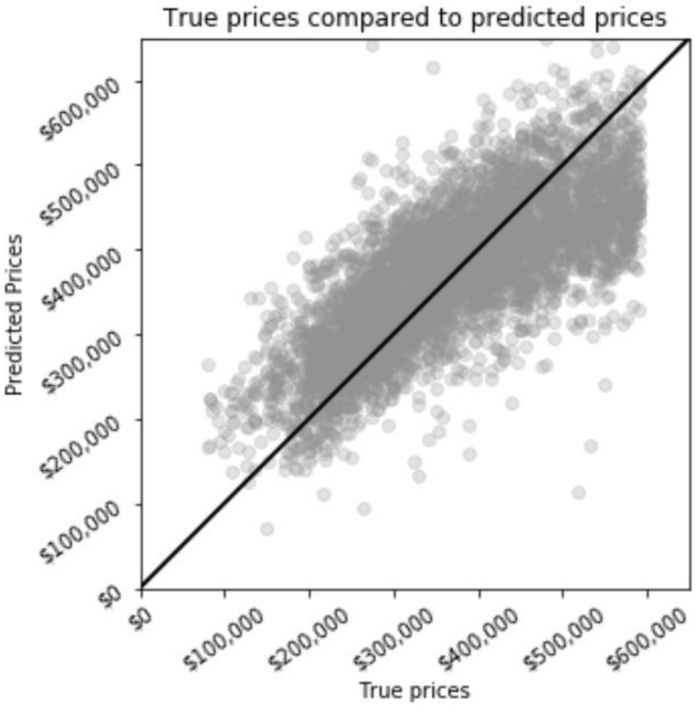
Cluster 2 Model



```
                            Features  Estimated Coefficients
0                           bedrooms           -1.229909e+04
1             grade_bucket_price_sqrt           3.231988e+02
2    sqft_living_bucket_price_squared           7.410774e-07
3                          bathrooms            2.437804e+04
4                         waterfront            1.295705e+05
5                          condition            1.240750e+04
6                              grade            3.218932e+04
7                           yr_built           -8.241704e+02
8                            zipcode           -4.188423e+01
9                                lat            4.057994e+05
10                              long            3.945118e+04
11                       sqft_living            1.685261e+01

Estimated intercept coefficient: -8974715.60852908

Summary Statistics
R-squared value: 0.5664798337761238
Root Mean Squared Error: 75550.38559401885
Mean Absolute Percentage Error (MAPE): 17.925521103105808
```



Clustering and Feature Encoding

9.28% improvement

	MAPE
All Features - no clustering	**25.84%**
Cluster 0	18.80%
Cluster 1	12.94%
Cluster 2	17.93%
Average of clusters	**16.56%**

Conclusion

It should be noted that this model is only trained on home sales in King County, Washington between 2014 and 2015. It’s use is limited to the same geographical area and time bounds, and should not be used to make predictions of home prices far into the future.