Modeling Real Estate by Predicting Residential Property Prices

Nihel Charfi, Olga Kozlova, James Reilly, James Wang

How are residential property prices calculated?

- For sale by owner
- Tax-assessed value
- Official appraisal
- Predictive models (public like Zillow and private like Realist)

Value Proposition

1. Identify undervalued properties
2. Identify opportunities to increase property value

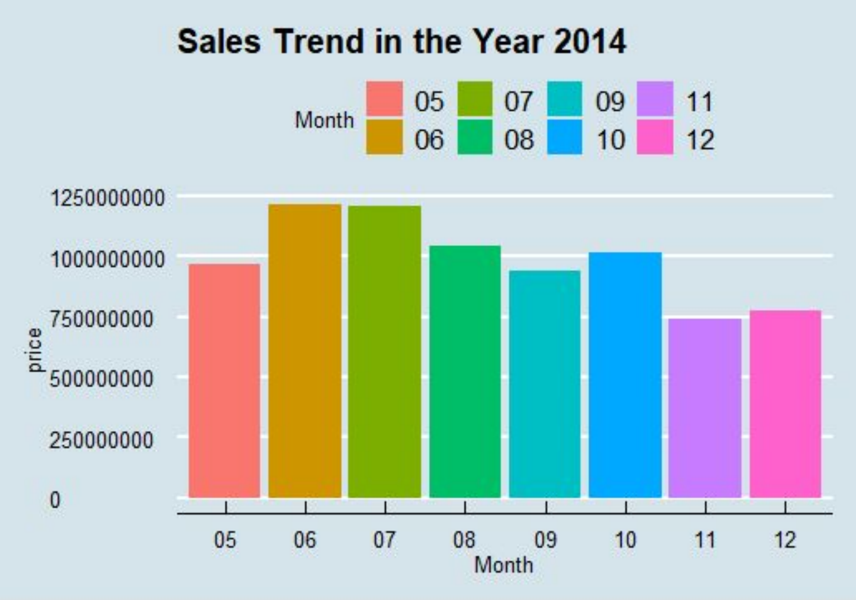
Dataset Description

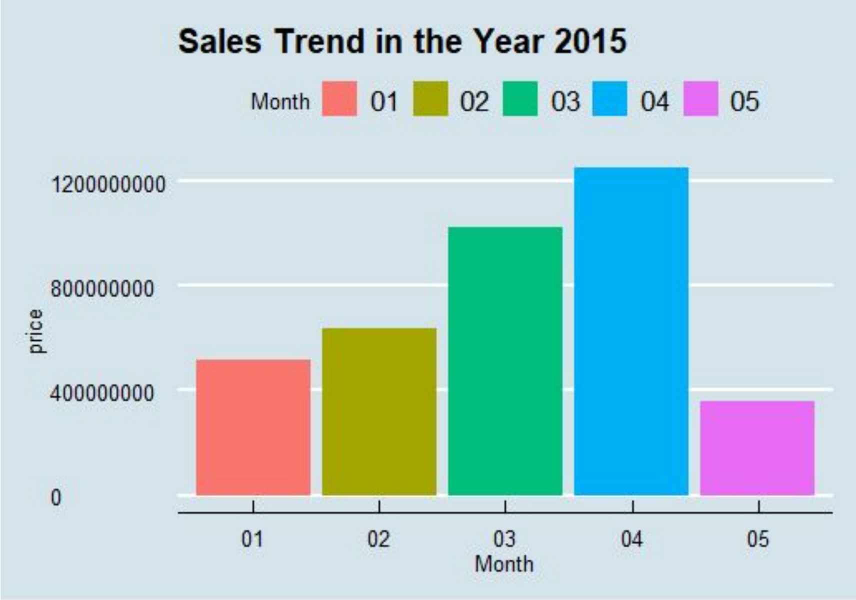
- 21,613 houses in King County (Washington State) sold between 2014-2015
- Physical properties about the homes
  - # Beds/Baths
  - Sqft. Living Space & Lot
  - Waterfront, View booleans
  - Condition & Grade (assessed by the King Country Grading System) factors
  - Year Built & Renovated (if applies)
  - Location Data (Zipcode, Latitude, Longitude)

Dataset Description

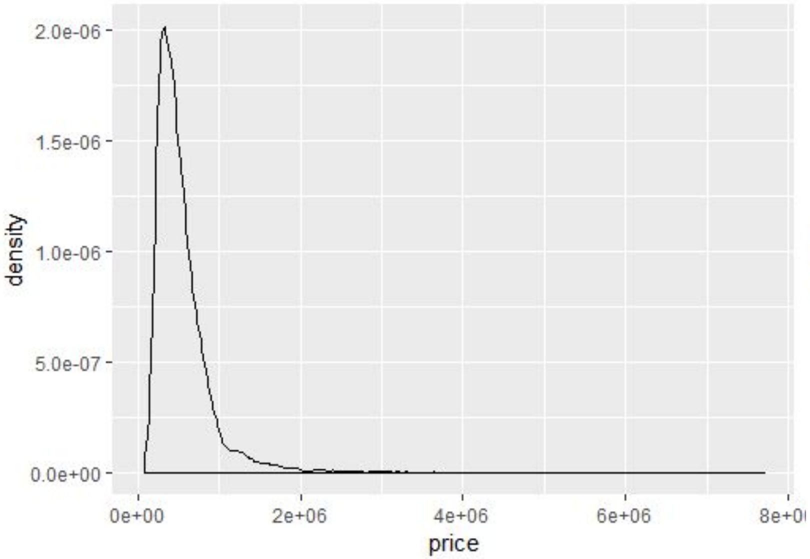
```
> str(hprice.df)
'data.frame':   21613 obs. of  22 variables:
 $ id           : num  7.13e+09 6.41e+09 5.63e+09 2.49e+09 1.95e+09 ...
 $ date         : Factor w/ 372 levels "20140502T000000",..: 165 221 291 221 284 11 57
..
 $ price        : num  221900 538000 180000 604000 510000 ...
 $ bedrooms     : int  3 3 2 4 3 4 3 3 3 3 ...
 $ bathrooms    : num  1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...
 $ sqft_living  : int  1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...
 $ sqft_lot     : int  5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...
 $ floors       : num  1 2 1 1 1 1 2 1 1 2 ...
 $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ view         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ condition    : int  3 3 3 5 3 3 3 3 3 3 ...
 $ grade        : int  7 7 6 7 8 11 7 7 7 7 ...
 $ sqft_above   : int  1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...
 $ sqft_basement: int  0 400 0 910 0 1530 0 0 730 0 ...
 $ yr_built     : int  1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...
 $ yr_renovated : int  0 1991 0 0 0 0 0 0 0 0 ...
 $ zipcode      : int  98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...
 $ lat          : num  47.5 47.7 47.7 47.5 47.6 ...
 $ long         : num  -122 -122 -122 -122 -122 ...
 $ sqft_living15: int  1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...
 $ sqft_lot15   : int  5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...
```

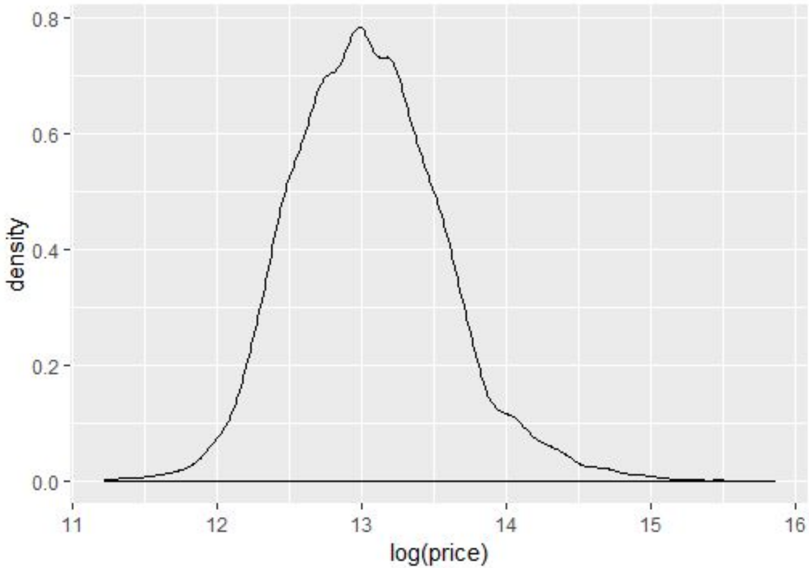
Sales trend from May 2014 to May 2015



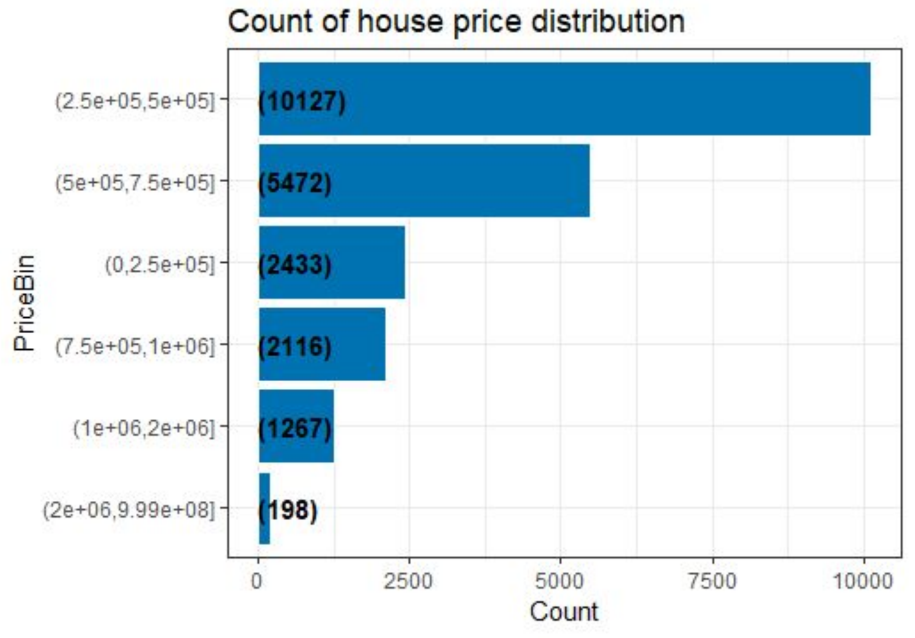


Price distribution in 2014 and 2015

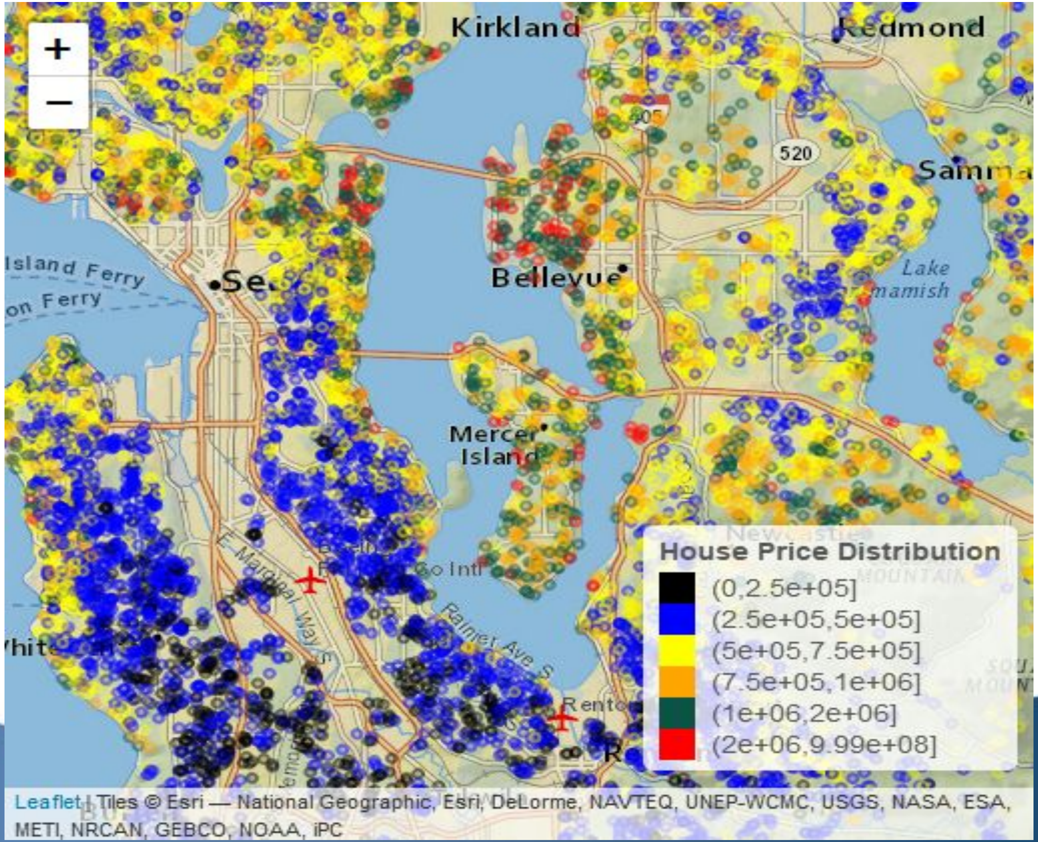




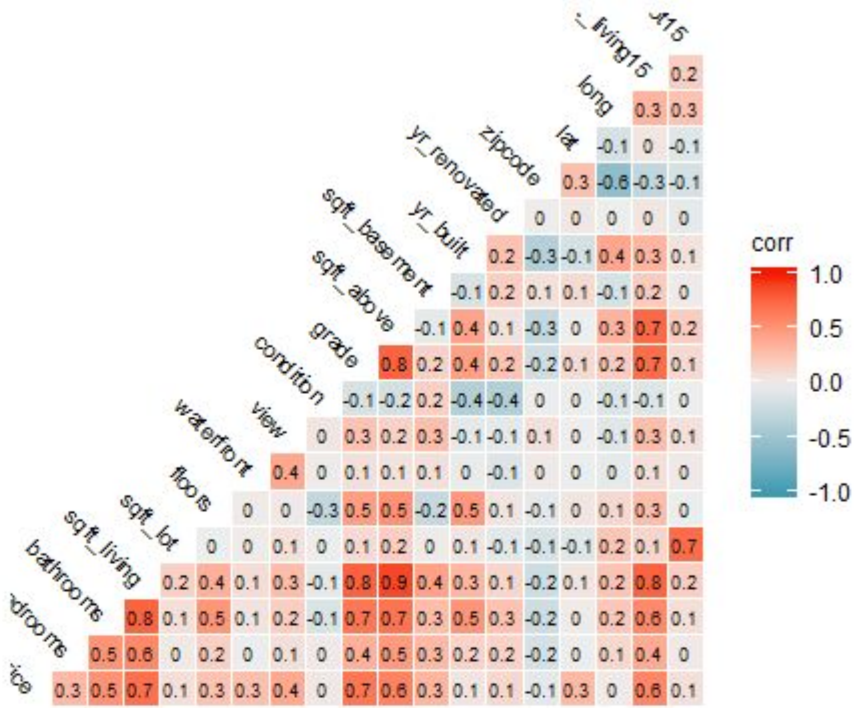
Price distribution in 2014 and 2015



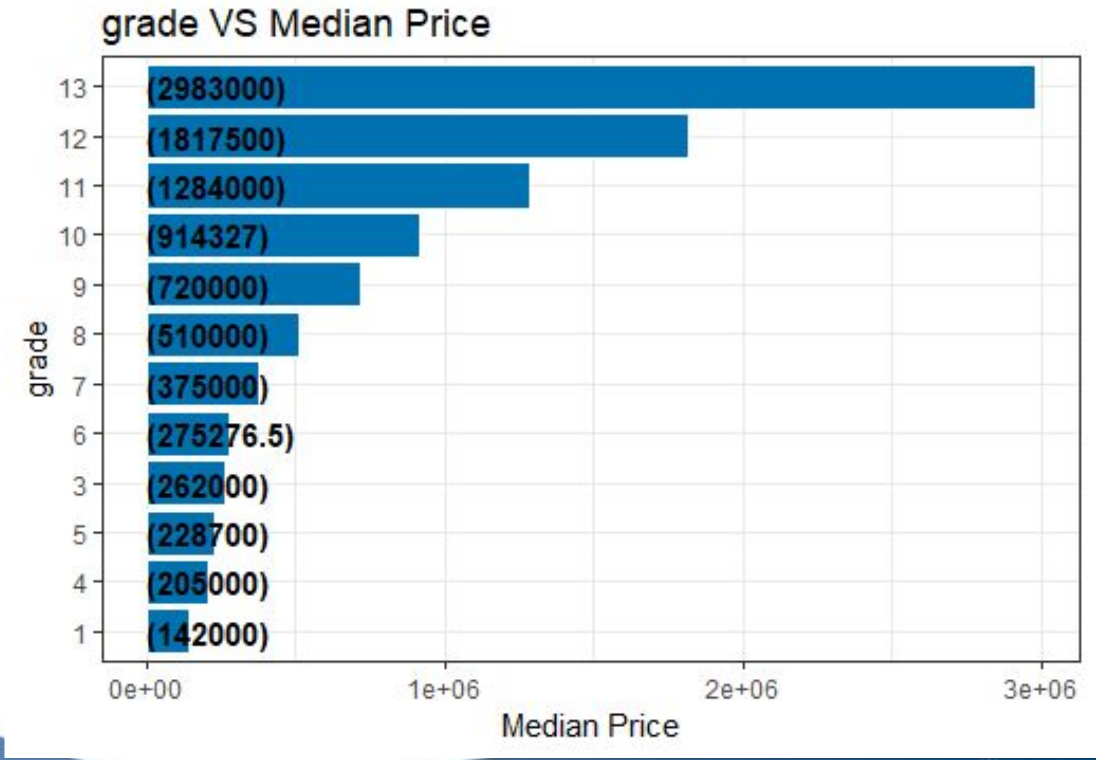
Houses location Map



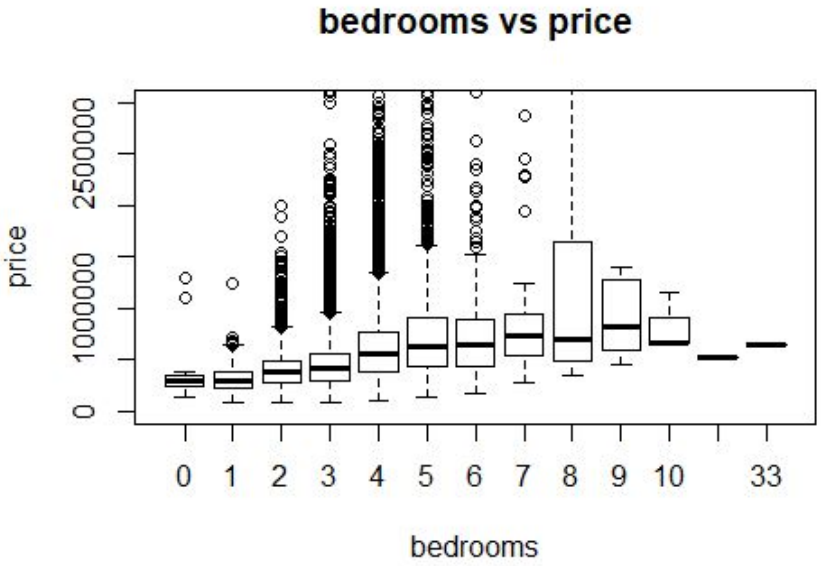
Dataset Visualization - Correlation

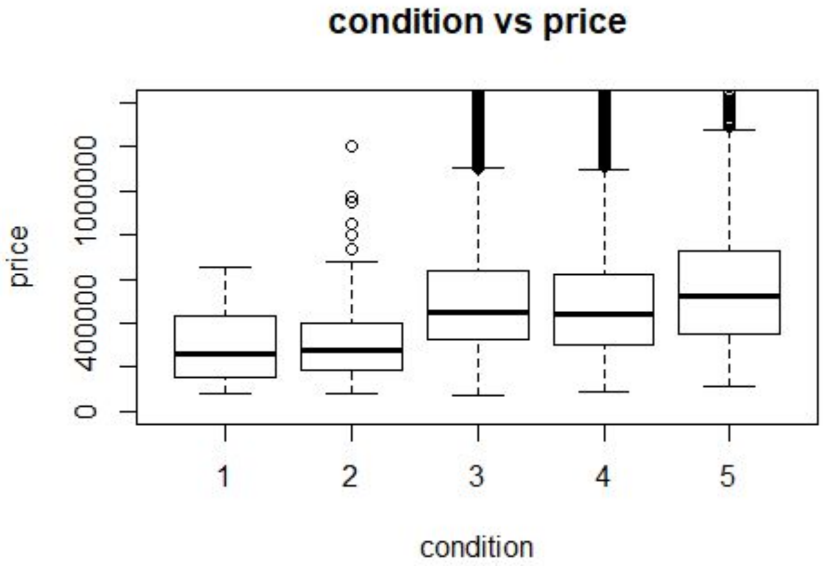


Dataset Visualization

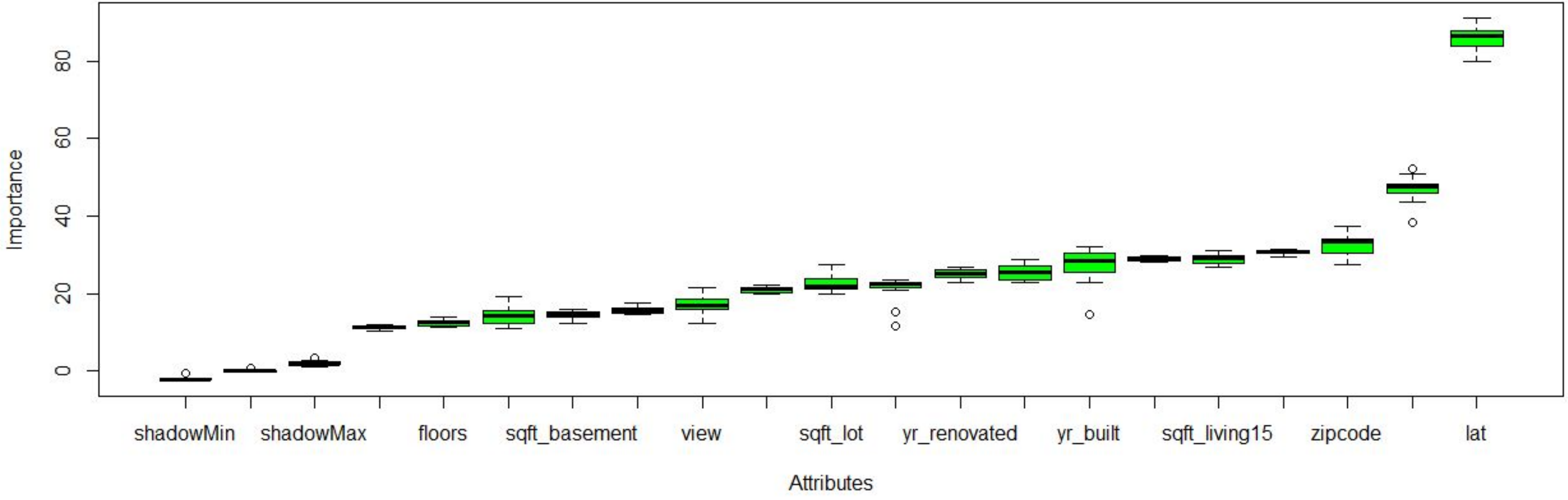


Dataset Visualization





Dataset Visualization - Boruta Feature Selection

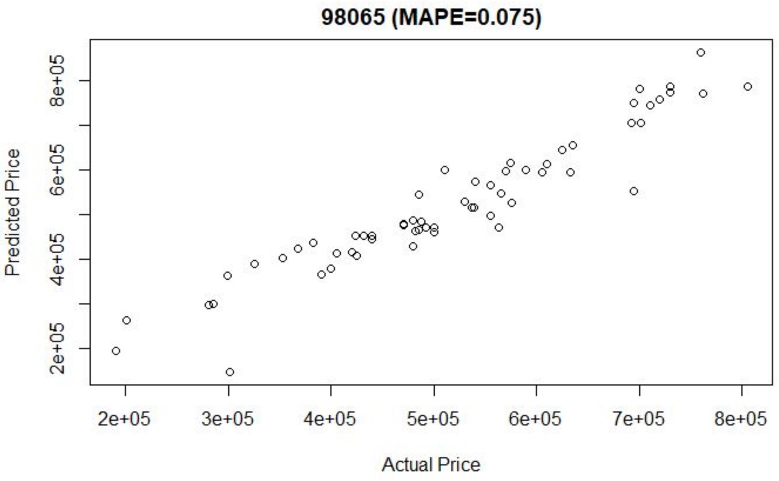


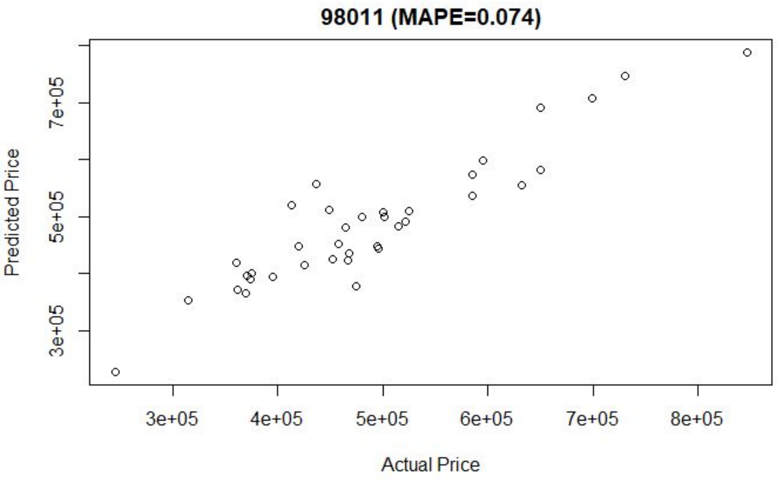
Comparison of Models

Model	Mean Absolute % Error
Multi-variate Linear Regression	21.192%
Random Forest	19.538%

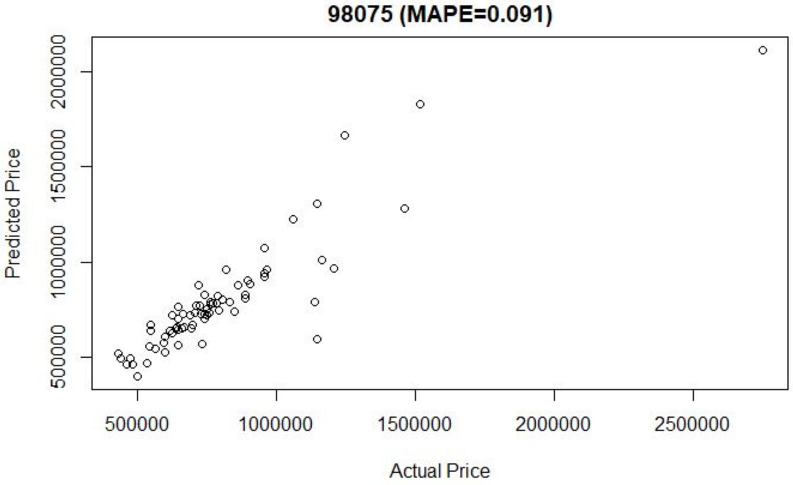
	numberOfHouses	zipcode	medianPrice	mape
63	195	98011	470000.0	0.07366011
11	141	98007	555000.0	0.08079399
64	274	98031	288200.0	0.08332194
56	321	98029	575000.0	0.08563075
55	310	98065	500000.0	0.08591979
43	359	98075	739999.0	0.09129048
21	256	98030	282255.0	0.09147443
20	351	98092	309780.0	0.09353712
17	199	98002	235000.0	0.09624872
3	283	98028	445000.0	0.10123318
6	405	98053	635000.0	0.10128134
5	441	98074	642000.0	0.10219683
10	590	98038	342000.0	0.10445080
15	190	98019	401250.0	0.10664537
27	455	98058	335000.0	0.11035906
34	229	98105	675000.0	0.11079546
59	234	98022	279500.0	0.11264953
33	57	98148	278000.0	0.11449169
35	548	98042	292000.0	0.11569475
48	125	98032	249000.0	0.11685478
18	494	98133	375000.0	0.11912783
19	282	98040	993750.0	0.15800295
61	446	98155	375000.0	0.15891846
23	269	98112	915000.0	0.15915241
40	317	98004	1150000.0	0.15942827
32	118	98070	463750.0	0.15967694
2	410	98125	425000.0	0.16242839
58	109	98109	736000.0	0.16364033
39	343	98144	450000.0	0.16946033
53	269	98168	235000.0	0.17342132
8	280	98198	265000.0	0.17465579
9	288	98146	305000.0	0.17733070
57	498	98006	760184.5	0.17838094
69	268	98055	294950.0	0.18167319
67	136	98188	264000.0	0.18356069
50	105	98102	720000.0	0.19291454
30	254	98166	390000.0	0.19821252
38	290	98122	572000.0	0.20161103
54	255	98177	554000.0	0.20290650
46	508	98118	367500.0	0.20292465
70	50	98039	1892500.0	0.20394688
68	124	98014	415000.0	0.20479676
62	81	98024	460000.0	0.20624120
22	184	98119	744975.0	0.21449782
45	100	98010	359999.5	0.23035717

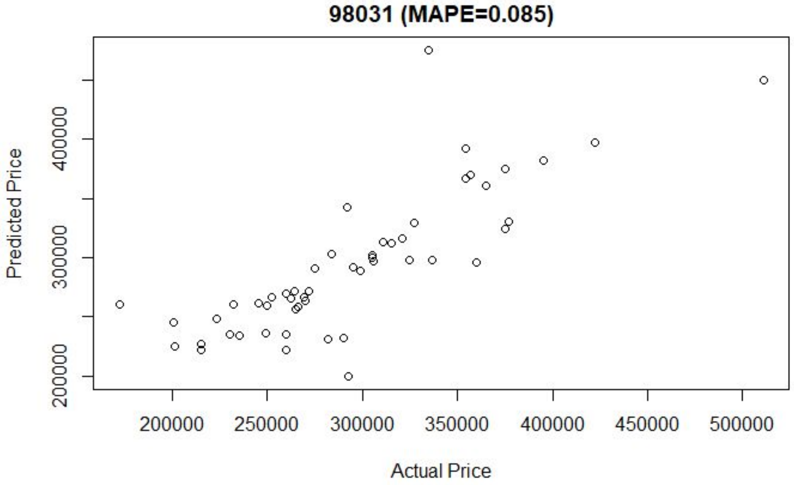
Best Predicted Zip Codes





Works on High & Low Value Homes





Potential Value in Purchase

- Our model shows that houses have sold below the predicted value
- The difference between our predicted price and the actual list price could be used to identify unrealized value
- The sum of $ in our best predicted Zip Code (98011) alone is $650,857.39

Summary of Potential Values

- In our best predicted Zip Codes the potential value is several million dollars in real estate

	zipcode	potentialValue
63	98011	650875.3923
56	98029	1019592.4962
24	98052	3242726.5204
6	98053	2555147.7926
55	98065	1392058.7915
64	98031	585143.4562
21	98030	585342.3603
17	98002	397962.2161
5	98074	3863813.4041
35	98042	1905701.3597
66	98072	1494259.6215
11	98007	1024022.5294
43	98075	4529958.6631
31	98023	1745144.3642
3	98028	1592410.5536
10	98038	3225895.6846
4	98136	1770939.2377
26	98117	3816846.1176
67	98188	394570.1002
7	98003	1323621.9835

Renovation Suggestions

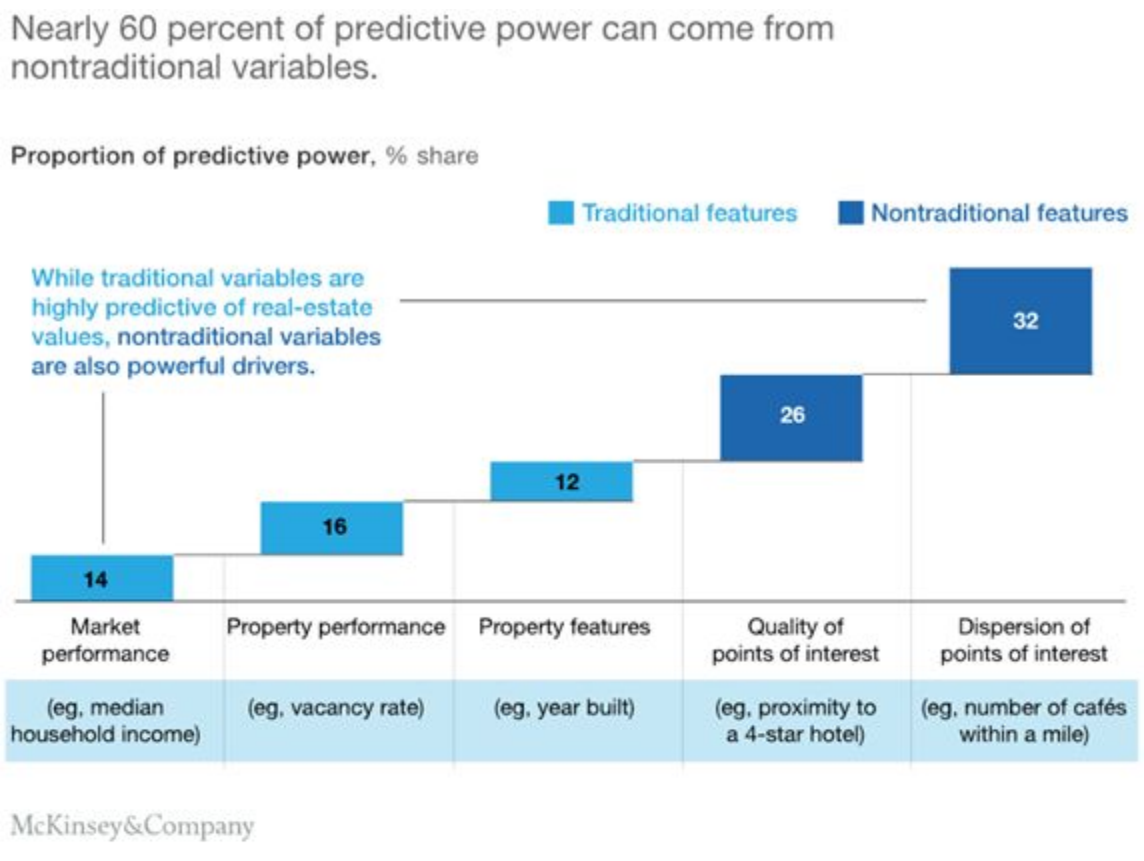
- LR Coefficients tell us the value of each variable in terms of price (Zip:98011)

Bedrooms	$554.09
Bathrooms	$13328.65
Sqft Living	$34.208
Sqft Basement	$65.85
Years Old	-$1594.40
Years Since Renovation	-$19.76
Sqft Above Ground	$104.29
Base Value	$238257.51

Conclusion

- Major assumption: consistency of resources within a zip code
- Location is the most important predictor
- Consistency of comparable sales matters (i.e., consistency of the market)

Possibilities for future research: Combining traditional (what sellers can control) and non-traditional data (what sellers cannot control)




Questions