Determinants of Housing Prices: An Econometric Analysis

Peter Finnerty

ISYS6640: Analytics and Business Intelligence

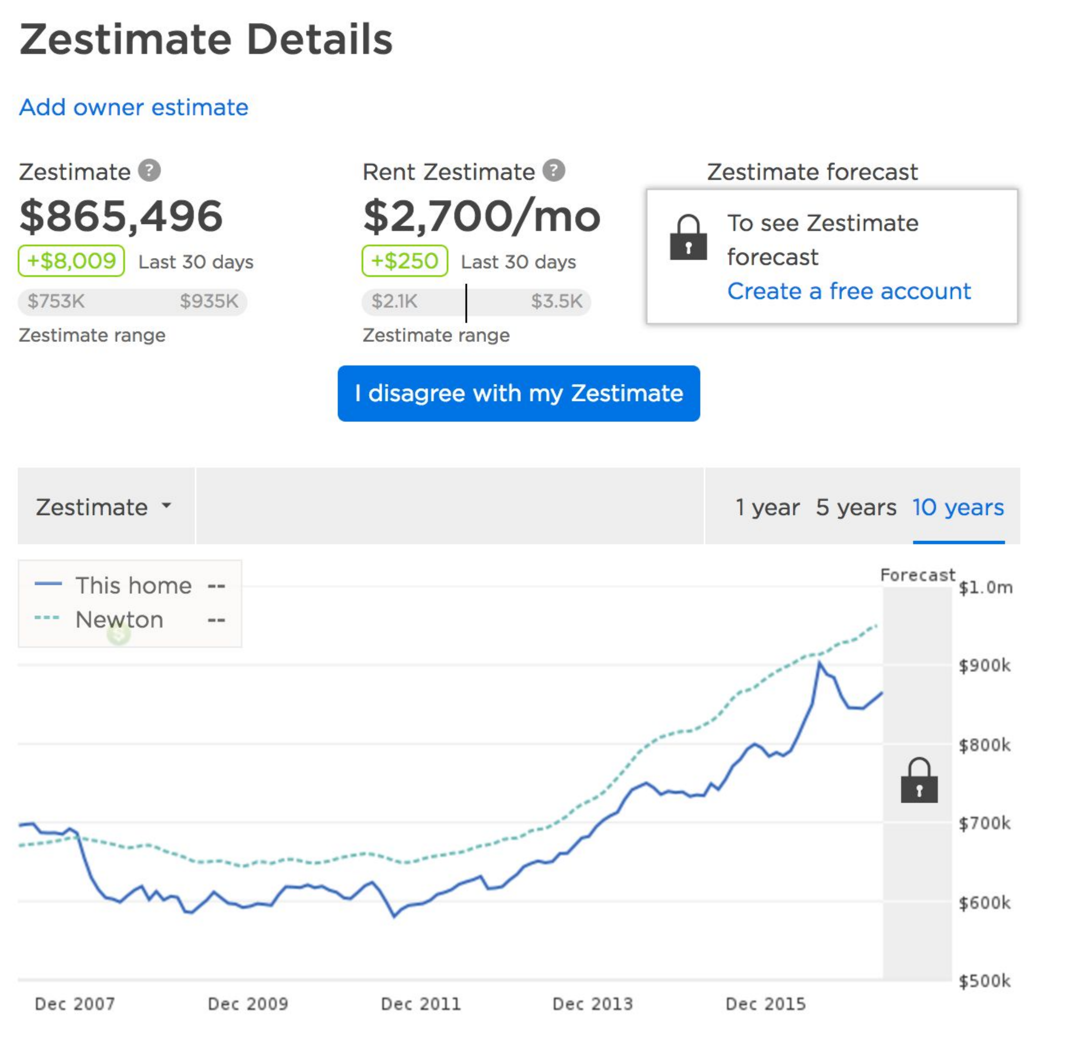
Agenda:

1. Background/Motivation
2. Data Overview
3. Feature Engineering
4. Models
5. Zillow Comparison
6. Conclusions

King County

Background: Home Pricing in the US

- Real Estate appraisal is a $6.5 billion industry
  - Appraisals are conducted by experts with knowledge of the area
- Buyers weigh a combination of factors in potential homes
  - Community: School, Neighborhood, Safety
  - Home features: Bedrooms, Bathrooms, Views, Renovations
- Zillow has become crucial to home buyers and sellers
  - Allows them to get a “Zestimate” of a home’s value
  - Quick assessment whether a home is over or under priced



Zestimate Details

Add owner estimate

Zestimate ?

$865,496

+$8,009 Last 30 days

$753K $935K

Zestimate range

Rent Zestimate ?

$2,700/mo

+$250 Last 30 days

$2.1K $3.5K

Zestimate range

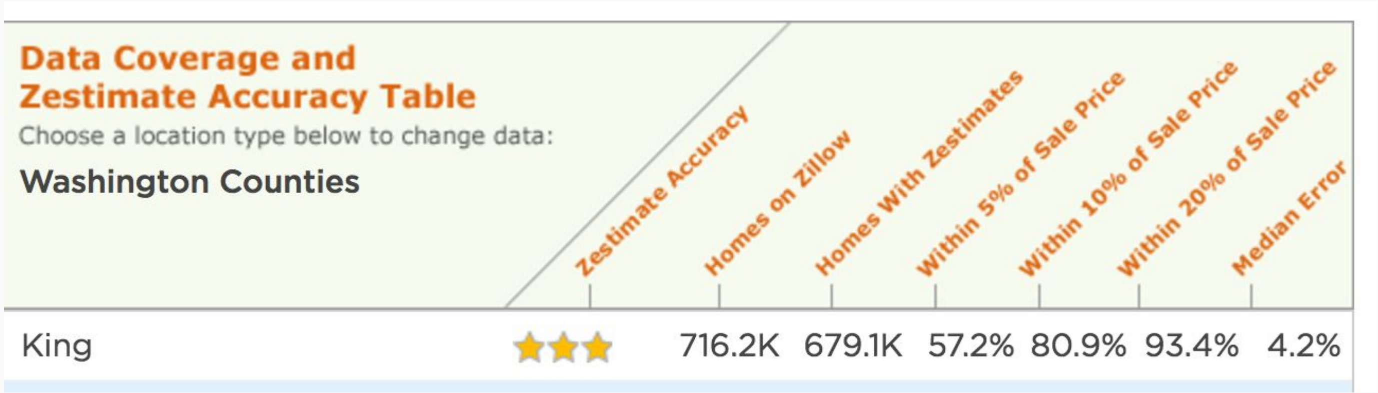
Zestimate forecast

To see Zestimate forecast

Create a free account

Our Motivation: How did Zillow calculate?

- What features of a home contribute to its housing price?
  - What should you consider when selling your home?
- Additionally, we wanted to see if we could beat the predictive accuracy of Zillow



Data Coverage and Zestimate Accuracy Table

Choose a location type below to change data:

Washington Counties

	Zestimate Accuracy	Homes on Zillow	Homes With Zestimates	Within 5% of Sale Price	Within 10% of Sale Price	Within 20% of Sale Price	Median Error
King	★★★	716.2K	679.1K	57.2%	80.9%	93.4%	4.2%

What did we find?

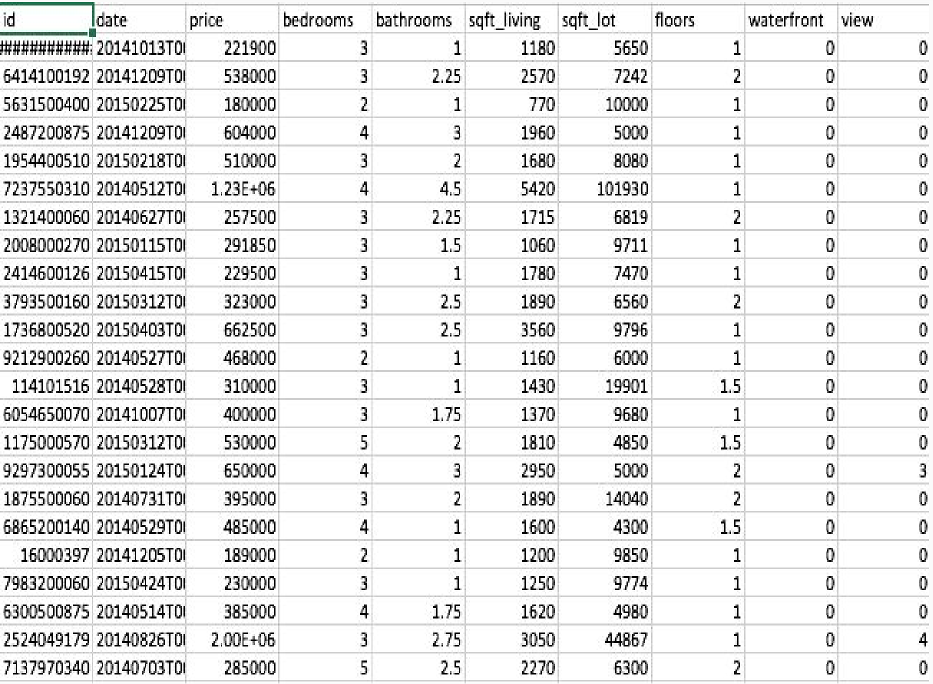
- Home prices are dependent on almost every factor
  - Bathrooms were our only insignificant variable
- Home prices are mainly impacted by what you expect:
  - Square Footage
  - Location
  - Waterfront
  - Quality of finishing (Grade)
- Prediction of home prices is incredibly difficult
  - Our predictive models were heavily skewed towards true and false negatives
- Zillow is fairly accurate and comprehensive

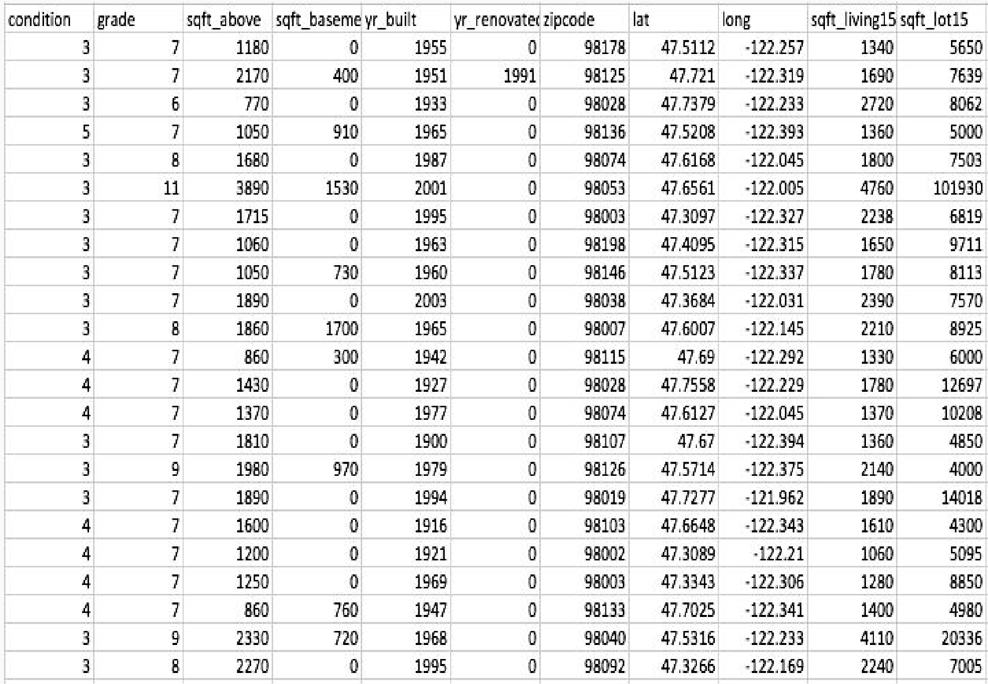
Data Overview

Our Data

- Homes sold in King County, WA from May 2014 to May 2015
  - Contains Seattle
- 19 different features
  - Including:
    - Price, Bedrooms, Bathrooms, Sqft_living, Floors, Waterfront
- 21,613 observations

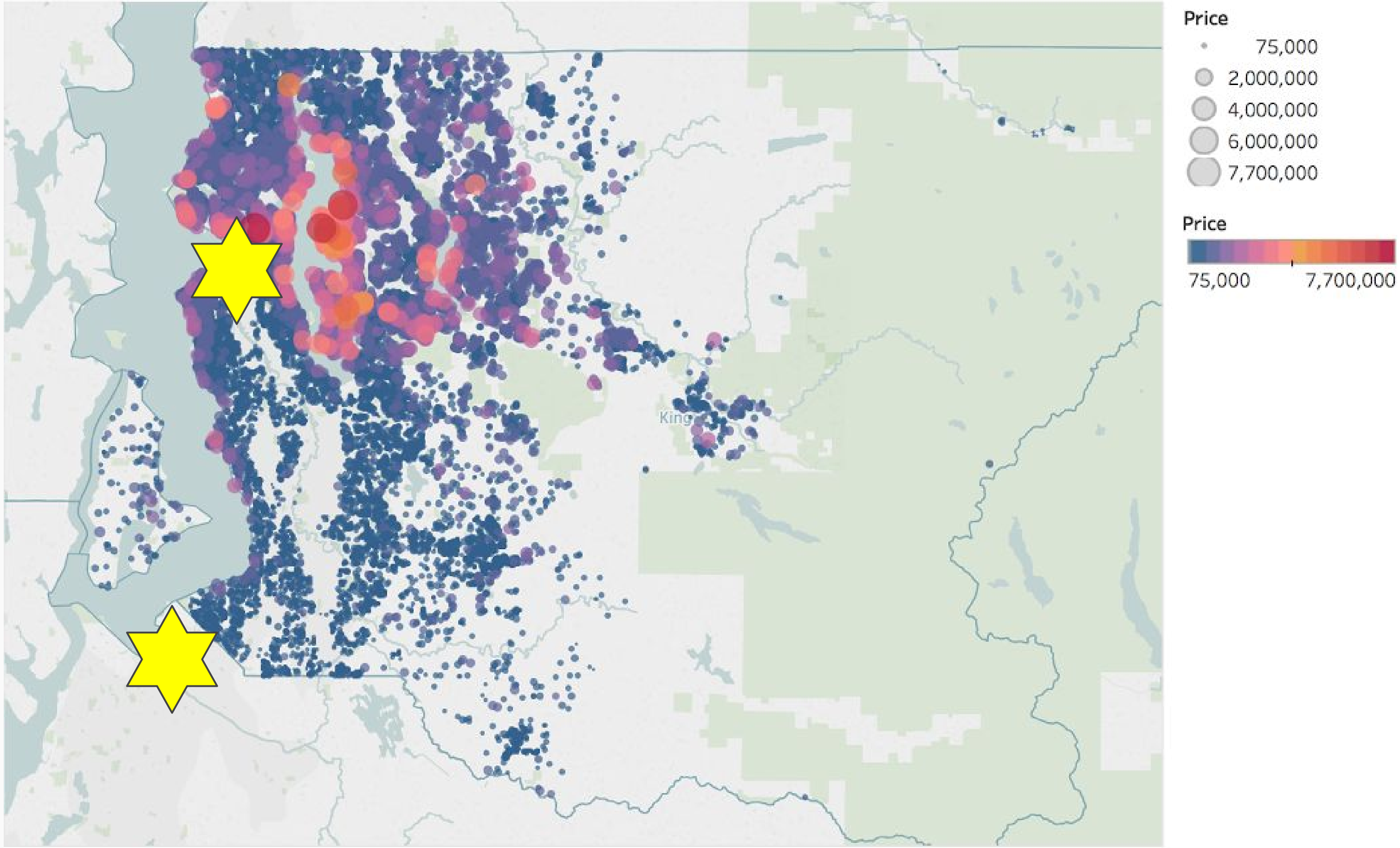
Overview of Data





id	date	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_baseme	yr_built	yr_renovatec	zipcode	lat	long	sqft_living15	sqft_lot15
###########:	20141013T0	221900	3	1	1180	5650	1	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
6414100192	20141209T0	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400	1951	1991	98125	47.721	-122.319	1690	7639
5631500400	20150225T0	180000	2	1	770	10000	1	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
2487200875	20141209T0	604000	4	3	1960	5000	1	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
1954400510	20150218T0	510000	3	2	1680	8080	1	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503
7237550310	20140512T0	1.23E+06	4	4.5	5420	101930	1	0	0	3	11	3890	1530	2001	0	98053	47.6561	-122.005	4760	101930
1321400060	20140627T0	257500	3	2.25	1715	6819	2	0	0	3	7	1715	0	1995	0	98003	47.3097	-122.327	2238	6819
2008000270	20150115T0	291850	3	1.5	1060	9711	1	0	0	3	7	1060	0	1963	0	98198	47.4095	-122.315	1650	9711
2414600126	20150415T0	229500	3	1	1780	7470	1	0	0	3	7	1050	730	1960	0	98146	47.5123	-122.337	1780	8113
3793500160	20150312T0	323000	3	2.5	1890	6560	2	0	0	3	7	1890	0	2003	0	98038	47.3684	-122.031	2390	7570
1736800520	20150403T0	662500	3	2.5	3560	9796	1	0	0	3	8	1860	1700	1965	0	98007	47.6007	-122.145	2210	8925
9212900260	20140527T0	468000	2	1	1160	6000	1	0	0	4	7	860	300	1942	0	98115	47.69	-122.292	1330	6000
114101516	20140528T0	310000	3	1	1430	19901	1.5	0	0	4	7	1430	0	1927	0	98028	47.7558	-122.229	1780	12697
6054650070	20141007T0	400000	3	1.75	1370	9680	1	0	0	4	7	1370	0	1977	0	98074	47.6127	-122.045	1370	10208
1175000570	20150312T0	530000	5	2	1810	4850	1.5	0	0	3	7	1810	0	1900	0	98107	47.67	-122.394	1360	4850
9297300055	20150124T0	650000	4	3	2950	5000	2	0	3	3	9	1980	970	1979	0	98126	47.5714	-122.375	2140	4000
1875500060	20140731T0	395000	3	2	1890	14040	2	0	0	3	7	1890	0	1994	0	98019	47.7277	-121.962	1890	14018
6865200140	20140529T0	485000	4	1	1600	4300	1.5	0	0	4	7	1600	0	1916	0	98103	47.6648	-122.343	1610	4300
16000397	20141205T0	189000	2	1	1200	9850	1	0	0	4	7	1200	0	1921	0	98002	47.3089	-122.21	1060	5095
7983200060	20150424T0	230000	3	1	1250	9774	1	0	0	4	7	1250	0	1969	0	98003	47.3343	-122.306	1280	8850
6300500875	20140514T0	385000	4	1.75	1620	4980	1	0	0	4	7	860	760	1947	0	98133	47.7025	-122.341	1400	4980
2524049179	20140826T0	2.00E+06	3	2.75	3050	44867	1	0	4	3	9	2330	720	1968	0	98040	47.5316	-122.233	4110	20336
7137970340	20140703T0	285000	5	2.5	2270	6300	2	0	0	3	8	2270	0	1995	0	98092	47.3266	-122.169	2240	7005

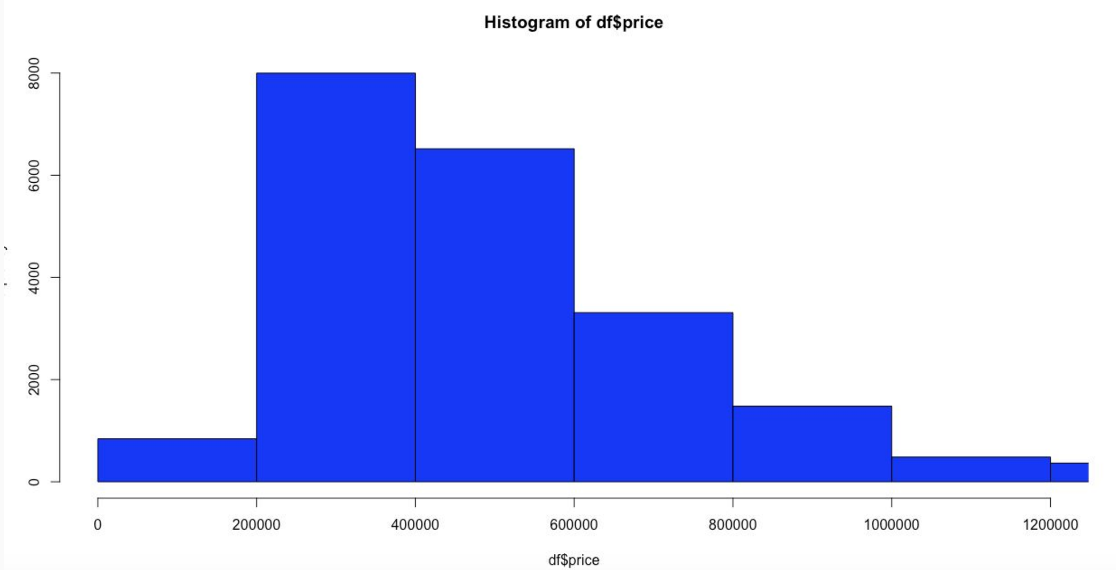
King County Home Sales



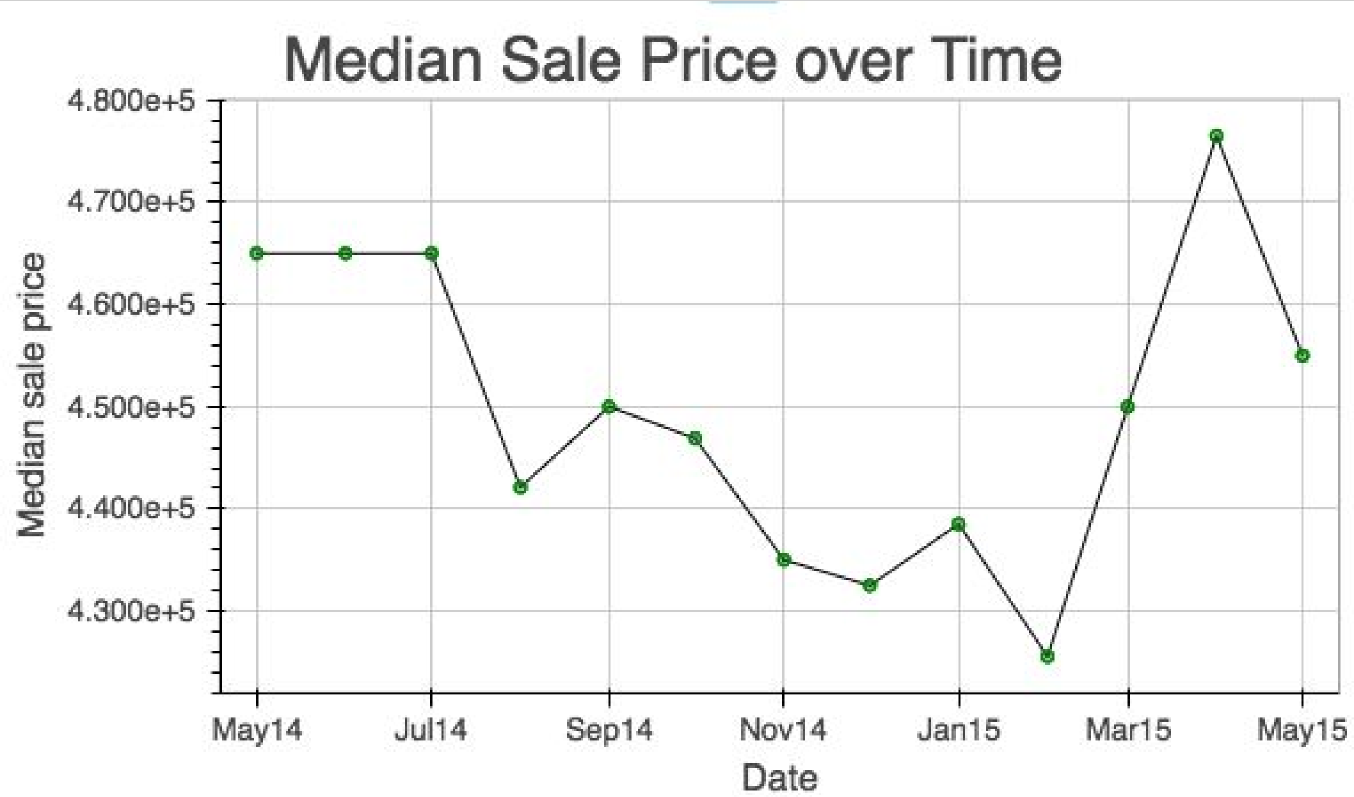
Map based on average of Longitude and average of Latitude. Color shows sum of Price. Size shows sum of Price. Details are shown for Id.

General Price Statistics

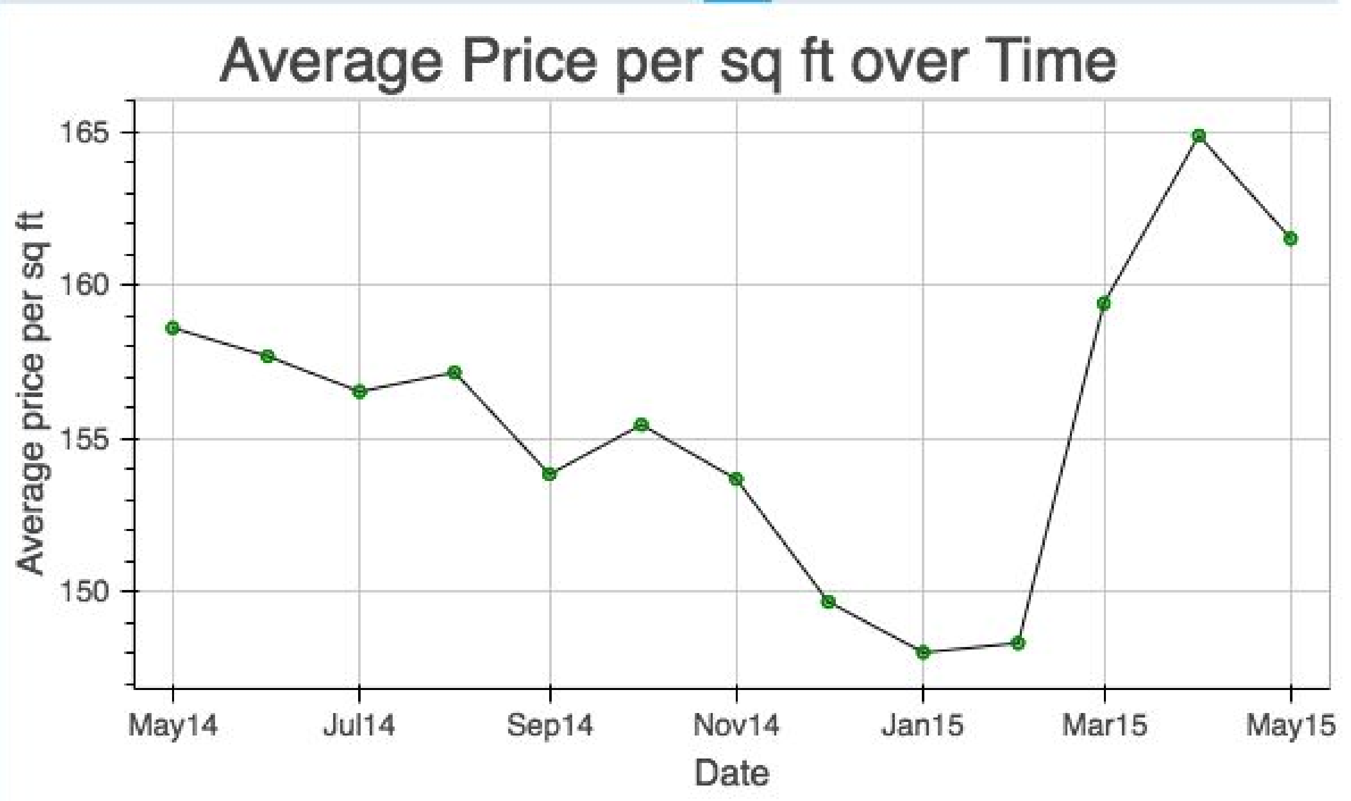
- Min: $75,000
- First Quartile: $322,000
- Median: $450,000
- Third Quartile: $645,000
- Max: $7,700,000
- Mean: $540,200



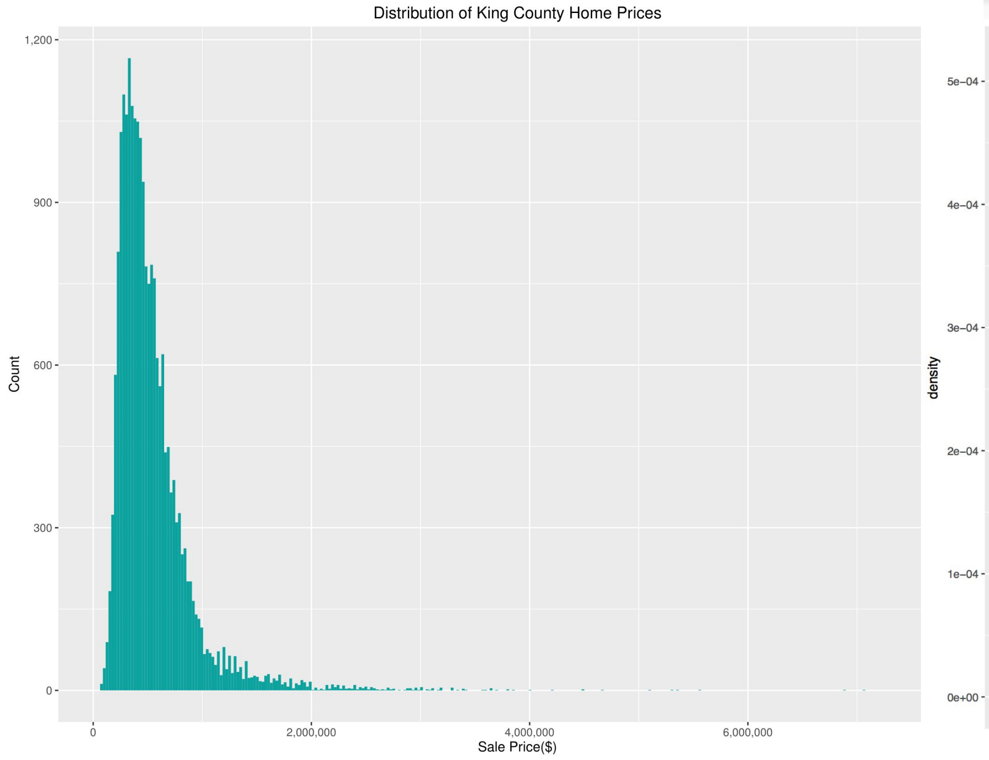
Median Sale Price over Time



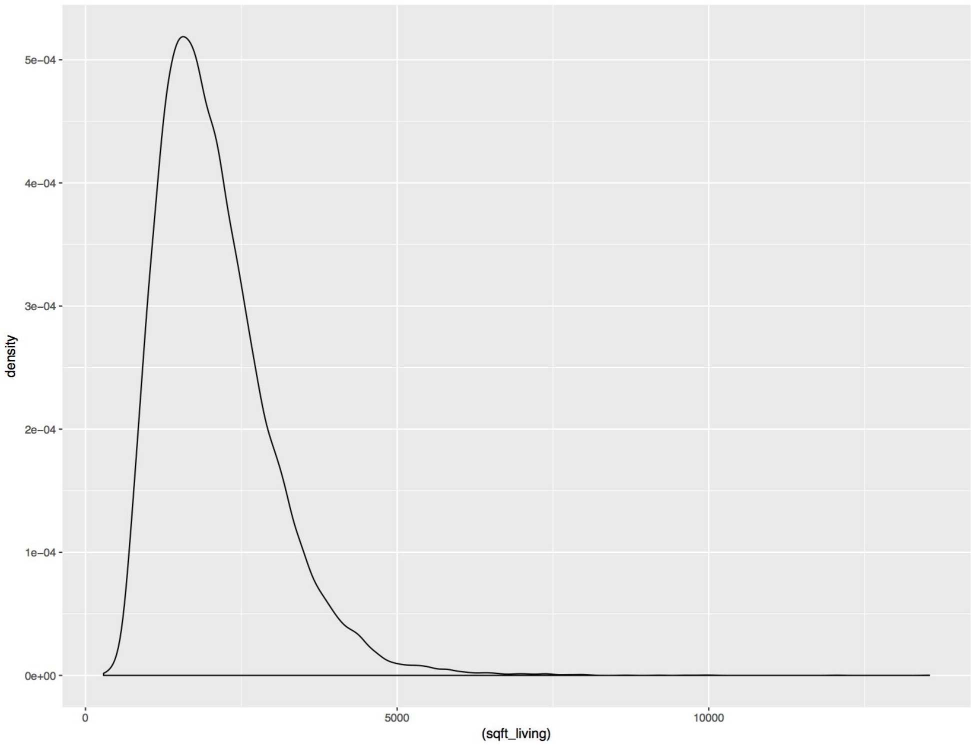
Distribution of Square Feet of Living Space



Distribution of Home Prices/Square Feet of Living Space

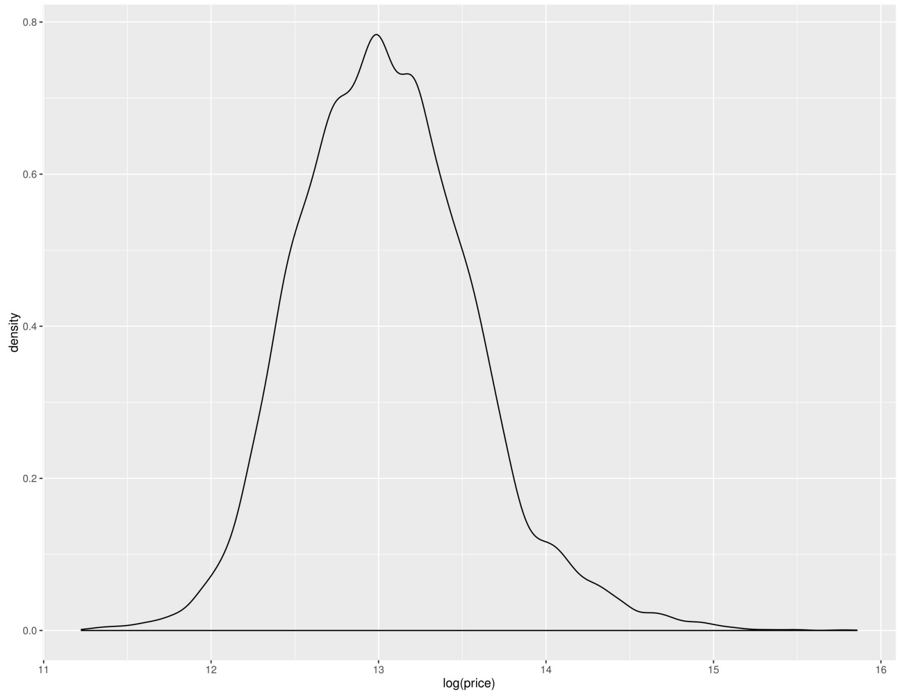


Price

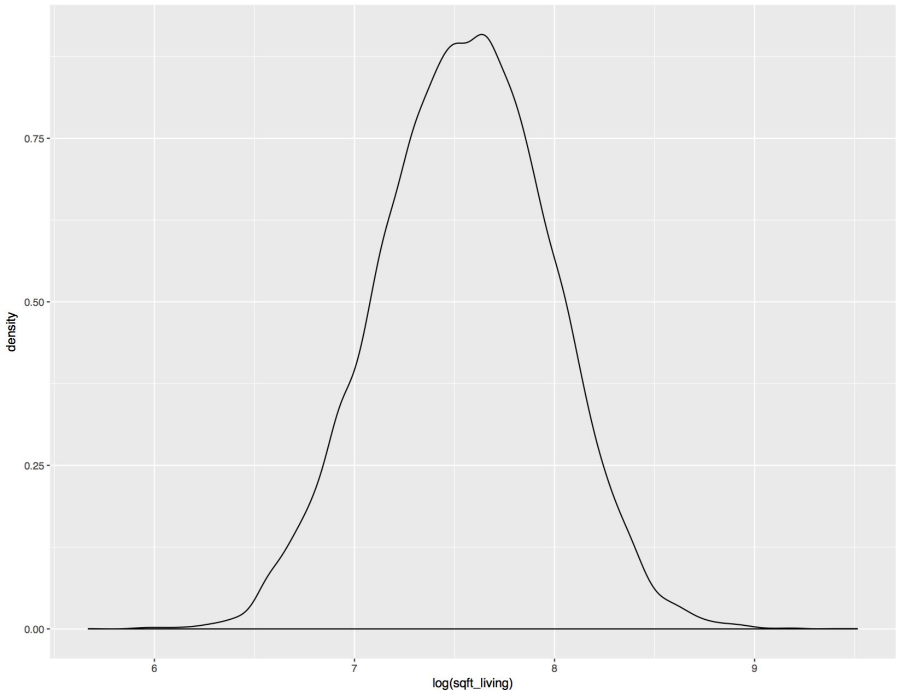


Sq. Feet of Living Space

Log Transformations

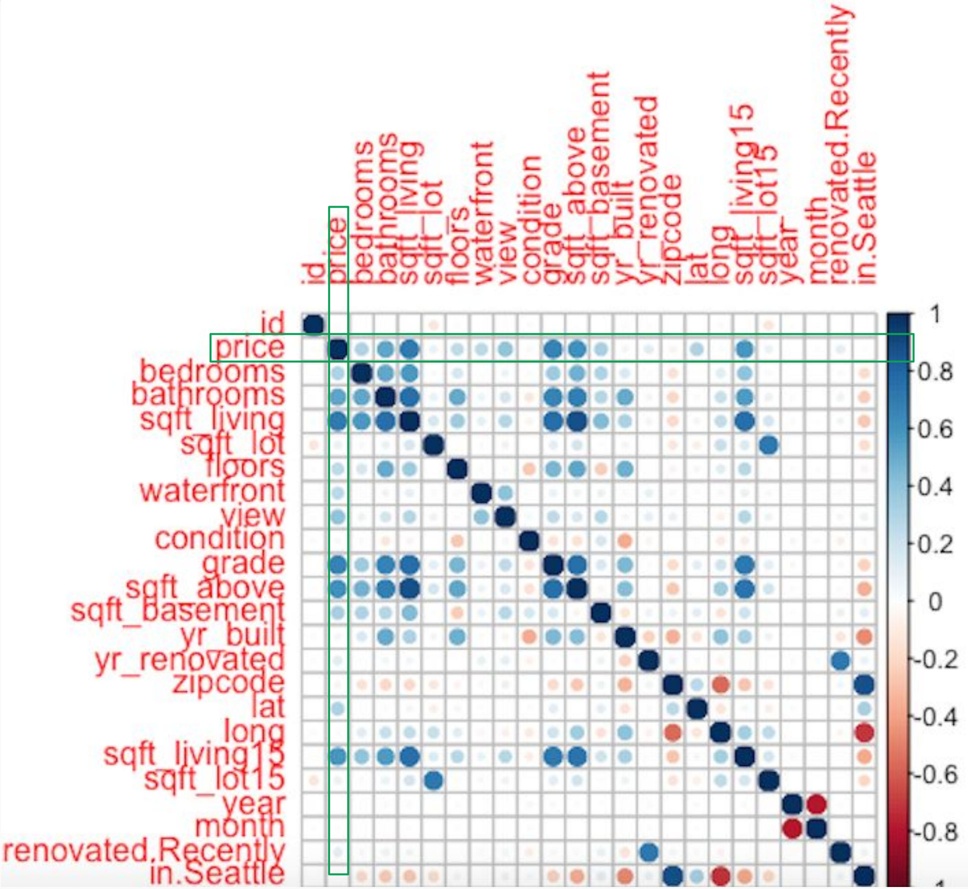


Price



Sq. Feet of Living Space

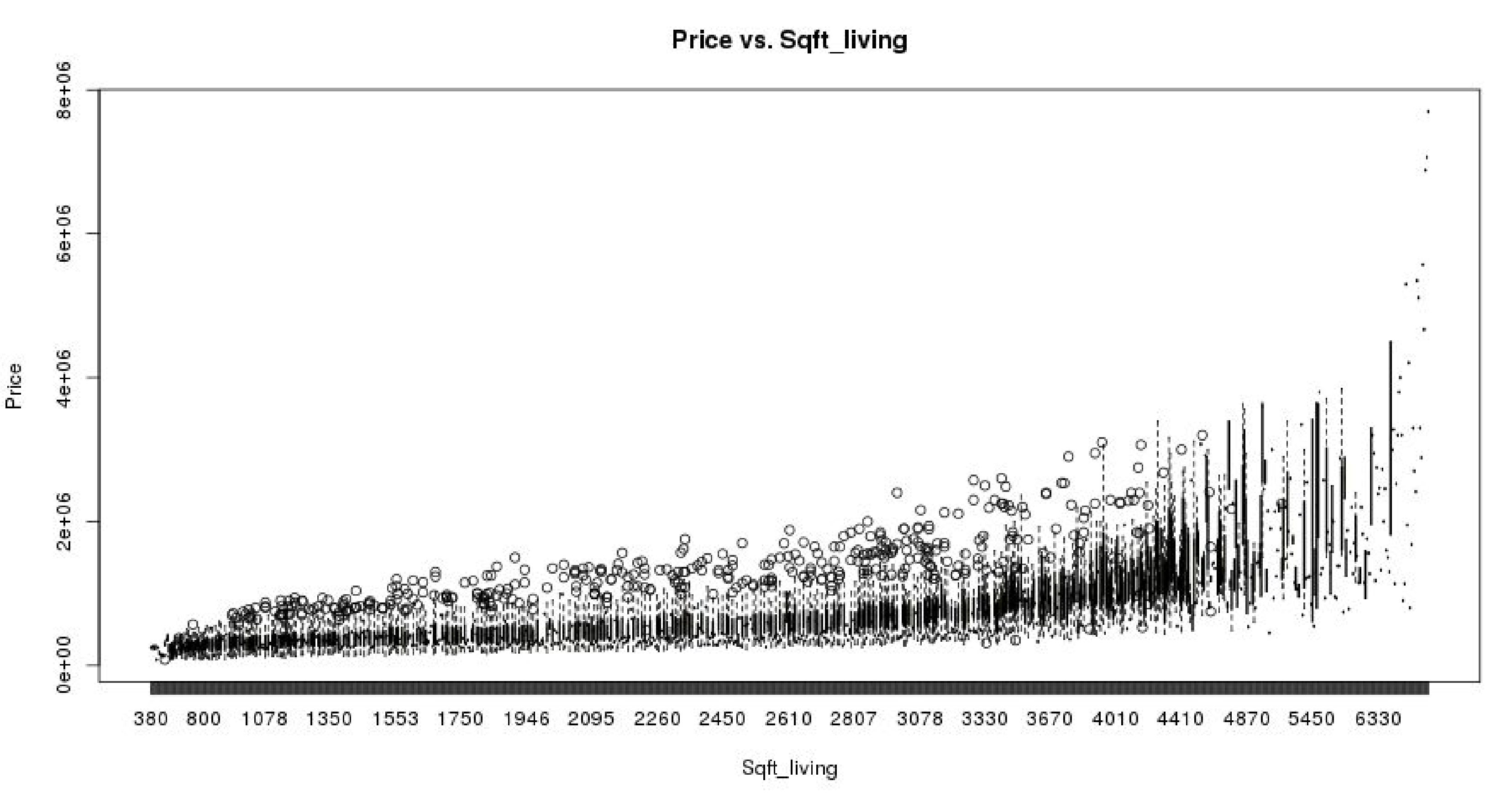
Correlogram



Top 5:

- Sqft_ living
- Grade
- View
- Sqft_above
- Sqft_basement

Price vs Square Foot Living Plot



Models

Feature Engineering

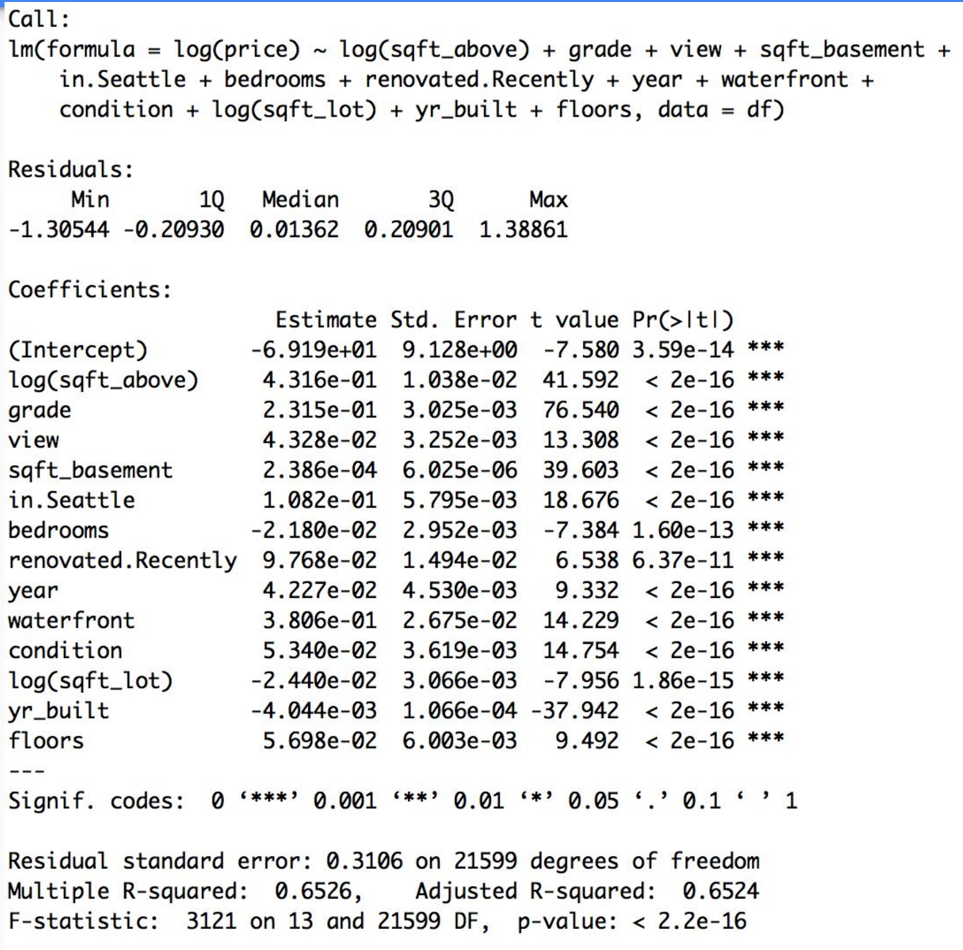
- in.Seattle: Binary
  - Whether or not a home’s zip code was in the city limits of Seattle
  - Really difficult to assess the value of being near a city
- renovated.Recently: Binary
  - Whether or not the home was renovated in the 15 years prior to sale
- Top_price_per_sqft
  - If a home’s price per square foot was in the top quarter of all homes sold
  - Divided price by square feet of living space and then filtered for top quartile
  - Used as our dependent variable for predictive modelling
- Year and Month
  - Were given in form of “20141013T000000”
  - Used lubridate library to separate into separate features

Model Selection

- Explanatory: OLS
  - Price is a continuous, unbound variable
  - Not measured by count
  - Goal is to explain what weighs most heavily on price
- Predictive: Logistic, Naive Bayes, and Decision Tree
  - Dependent (top 25% of price/sqft) was binary
  - Probability of a house being in top 25% of price/sqft
    - Probabilistic required Logistic, NB, or DT
  - Goal is to predict an expensive home

OLS Regression

- Multiple R-squared: .6526
- Variables are significant
- Most Surprising
  - Year built
  - Basement
  - View
- Logged:
  - Sqft_above
  - Sqft_lot
- Insignificant and Removed:
  - Bathrooms
  - Sqft_living



```
Call:
lm(formula = log(price) ~ log(sqft_above) + grade + view + sqft_basement + 
    in.Seattle + bedrooms + renovated.Recently + year + waterfront + 
    condition + log(sqft_lot) + yr_built + floors, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.30544 -0.20930  0.01362  0.20901  1.38861 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -6.919e+01  9.128e+00  -7.580 3.59e-14 ***
log(sqft_above)     4.316e-01  1.038e-02  41.592  < 2e-16 ***
grade               2.315e-01  3.025e-03  76.540  < 2e-16 ***
view                4.328e-02  3.252e-03  13.308  < 2e-16 ***
sqft_basement       2.386e-04  6.025e-06  39.603  < 2e-16 ***
in.Seattle          1.082e-01  5.795e-03  18.676  < 2e-16 ***
bedrooms           -2.180e-02  2.952e-03  -7.384 1.60e-13 ***
renovated.Recently  9.768e-02  1.494e-02   6.538 6.37e-11 ***
year                4.227e-02  4.530e-03   9.332  < 2e-16 ***
waterfront          3.806e-01  2.675e-02  14.229  < 2e-16 ***
condition           5.340e-02  3.619e-03  14.754  < 2e-16 ***
log(sqft_lot)      -2.440e-02  3.066e-03  -7.956 1.86e-15 ***
yr_built           -4.044e-03  1.066e-04 -37.942  < 2e-16 ***
floors              5.698e-02  6.003e-03   9.492  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3106 on 21599 degrees of freedom
Multiple R-squared:  0.6526,	Adjusted R-squared:  0.6524 
F-statistic:  3121 on 13 and 21599 DF,  p-value: < 2.2e-16
```

Coefficient Interpretation

So what does this tell us?

- Our most impactful attributes were waterfront (38%), grade(23%), a house in Seattle(11%), and a recently renovated house (10%), in that order.

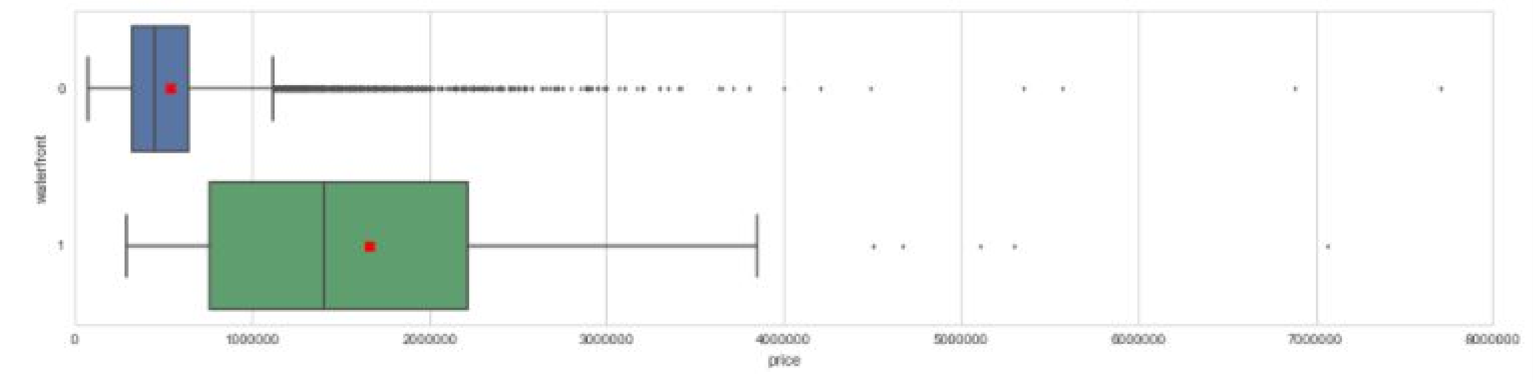
- Ex: A 10% increase in sqft_above will result in $1.10^{.4316}$= 4.20% Increase in price
  - Log independent, log-dependent variable

What did we expect?

- Since the county and major city of Seattle were so close to the water, we expected waterfront and in.Seattle to be the most impactful price raisers. We also intuitively expected recently renovated houses to yield a high impact.

- While we were fairly accurate in our guesses, we expected Seattle and the recently renovated houses to yield a higher impact than they did. We also didn’t expect the grade to be as impactful as it was, but it intuitively makes sense as to why it has some impact.

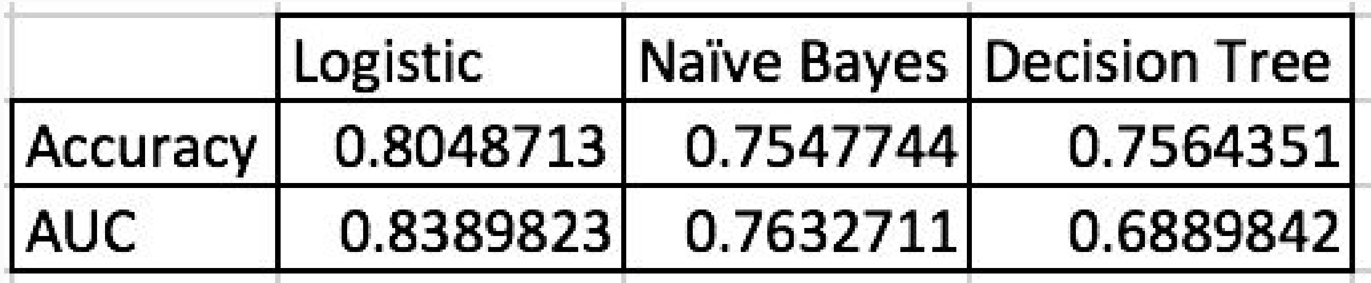
Impact of Waterfront



- The no waterfront boxplot is very short, indicating that the prices sold of these houses are very close together

- The waterfront boxplot is much longer, suggesting housing prices with a waterfront differ greatly
    - In general, waterfront houses sell for a higher price than non-waterfront ones, with its median being almost $1.5 million dollars

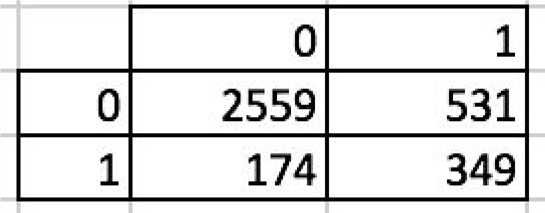
Predictive Models: Can we predict an expensive home?



	Logistic	Naïve Bayes	Decision Tree
Accuracy	0.8048713	0.7547744	0.7564351
AUC	0.8389823	0.7632711	0.6889842

- Dependent Variable: Top 25% Price per Square Foot
  - Created this categorical value for analysis
  - Wanted to eliminate the importance of square footage
  - Removed Sqft variables to maintain independence
- Split into Training and Test
  - Training: 1 to 18,000
  - Test: 18,001 to 21,613
- Best results from the Logistic Model
  - 80.49% chance that our model correctly predicts a home to be in the top 25%
  - Concern over false negatives
  - About 50% accuracy when considering distribution of home prices

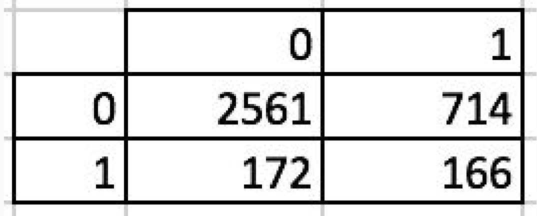
Logistic Confusion Matrix



	0	1
0	2559	531
1	174	349

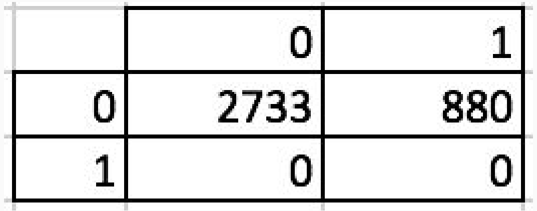
What about the other models?

Naive Bayes



	0	1
0	2561	714
1	172	166

Decision Tree



	0	1
0	2733	880
1	0	0

- Not as good as the Logistic Model
- We feel these are close to negative predictors
  - Especially true of the decision tree
- Show the dangers of looking just at accuracy and AUC scores

How do we compare to Zillow?

- Not too favorably
  - Median Percent Error (OLS): 21.978%
  - Zillow Median Error in King County: 4.2%
- Why are we different?
  - Zillow accounts for the community factors we could not control
  - Zillow has access to more than one year's worth of records
  - Zillow's scope in similar cities and locations
  - Input from owners, real estate agents, and consumers
- Are our results important?
  - Yes, they provide insight into how Zillow weighs factors of a home
  - We are skeptical of our prediction abilities

How Accurate is Your Zestimate?

Conclusions

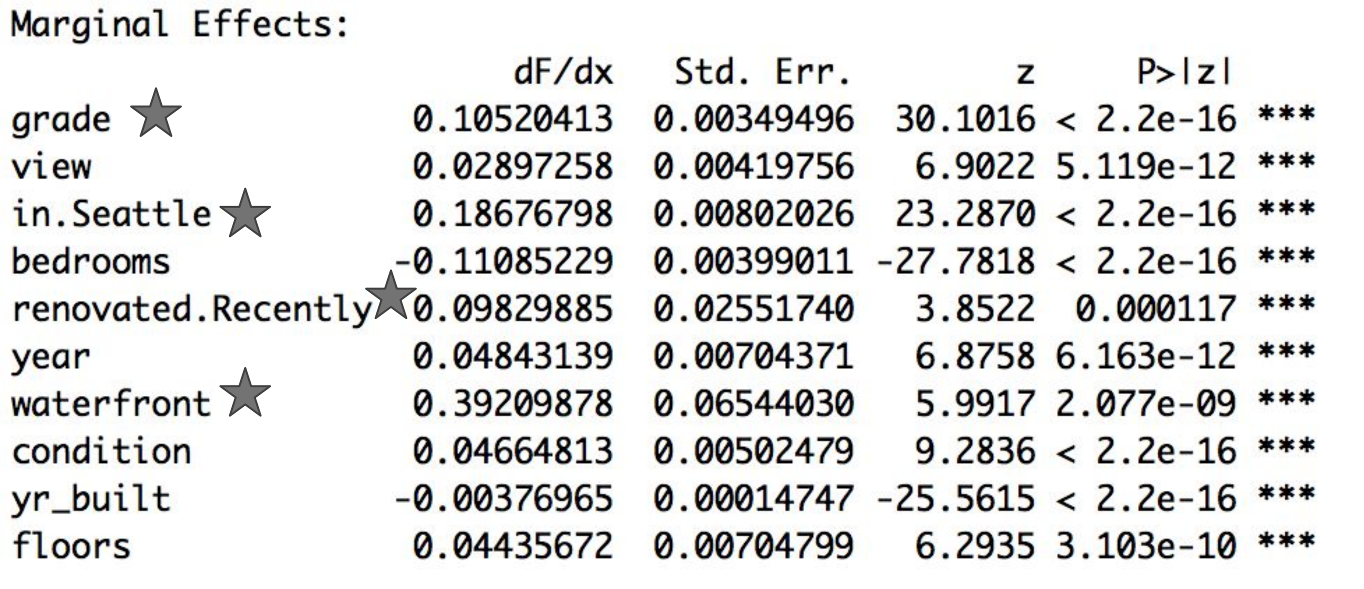
- In this project, we used OLS and Logistic models to understand home pricing
    - OLS: How does Zillow weigh home factors in estimating?
    - Logistic: What predicts an expensive home?
- Our takeaways:
    - Square footage, Grade, Location, and Renovations all significantly impact pricing
    - Must consider community features in addition to home features
- Predictive Models fail to account for the complexities of the home market
- Hire an appraiser, not a data analyst

Questions?

Appendix One: Tricky Code

- in.Seattle Feature:
    - df$in.Seattle <- ifelse((df$zipcode %in% seattlezipcodes$zip), 1, 0)
- recently.Renovated Feature:
    - Created with: df$renovated.Recently <- ifelse((df$year - df$yr_renovated) <= 15, 1, 0)
- Price/Sqft Feature
    - df$pricepersqft <- df$price/df$sqft_living
    - df$top_price_per_sqft <- ifelse(df$pricepersqft > 318.40, 1, 0)
    - df$top_price_per_sqft <- as.factor(df$top_price_per_sqft)

Appendix Two: What does our Logistic Model mean?



```
Marginal Effects:
                         dF/dx    Std. Err.        z      P>|z|
grade ★             0.10520413  0.00349496  30.1016 < 2.2e-16 ***
view                0.02897258  0.00419756   6.9022 5.119e-12 ***
in.Seattle ★        0.18676798  0.00802026  23.2870 < 2.2e-16 ***
bedrooms           -0.11085229  0.00399011 -27.7818 < 2.2e-16 ***
renovated.Recently★ 0.09829885  0.02551740   3.8522  0.000117 ***
year                0.04843139  0.00704371   6.8758 6.163e-12 ***
waterfront ★        0.39209878  0.06544030   5.9917 2.077e-09 ***
condition           0.04664813  0.00502479   9.2836 < 2.2e-16 ***
yr_built           -0.00376965  0.00014747 -25.5615 < 2.2e-16 ***
floors              0.04435672  0.00704799   6.2935 3.103e-10 ***
```

Change in predicted probability that home will have price/sqft in top 25%