STATISTICAL ANALYSIS OF HOUSING CHARACTERISTICS AND THEIR EFFECTS ON RESIDENTIAL HOUSE VALUES

Snehal Shetye

Thanh Pham

Shrilekha Singh

Seraph Shen

Faculty Advisor: Dr. Martina Bremer

King County Housing Data

- From Kaggle
- 19 housing features, along with ID and price
- 21613 Observations
- May 2014 - May 2015

Aim

- We are interested in how housing prices are affected by the houses' features and launch further investigations to explore possible relationships between them.

- The purpose of our statistical analysis is to predict housing sales prices in King County, WA.

16 Variables Selected

	Price($)	Bedrooms	Bathrooms	Living Area (sq ft)	Lot size (sq ft)	Floors	Waterfront	View
Median	450,000	3	2.25	1910	7618	1.5	-	0
Range	75,000-7,700,000	0-33	0.00-8.00	290 -13,540	520 -1,651,359	1.0-3.5	0<-21450 1<-163	0-4
	Condition	**Grade**	**Year built**	**Year renovated**	**Latitude (degree)**	**Longitude (degree)**	**Living Area 2015 (sq ft)**	**Lot Size 2015 (sq ft)**
Median	3	7	1975	0	47.56	-122.2	1987	7620
Range	1-5	4-13	1900-2015	0-2015	47.16 -47.78	(-122.5) - (-121.3)	399 -6,210	651 -871,200

5 Variables Removed

1. ID
2. Date
3. Zip code
4. Sqrt footage of the house apart from the basement
5. Sqrt footage of the basement

Irrelevant

Substituted

Linear combination of “living area”

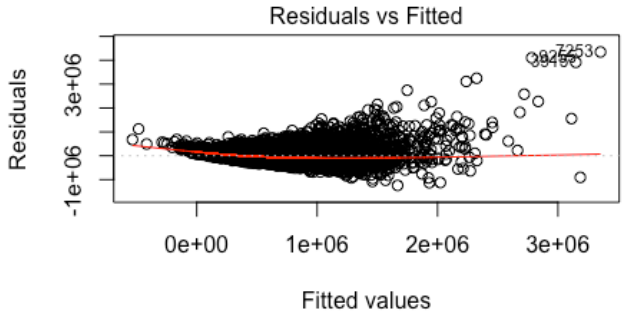
Sample Determination & Data Diagnostics

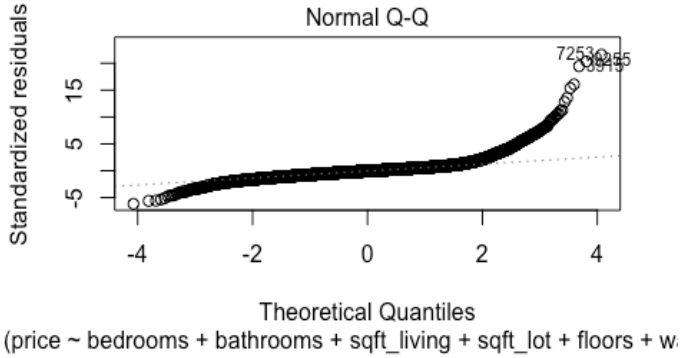
Data splitting -> Ratio: 70% - 30%

Fitting a raw model

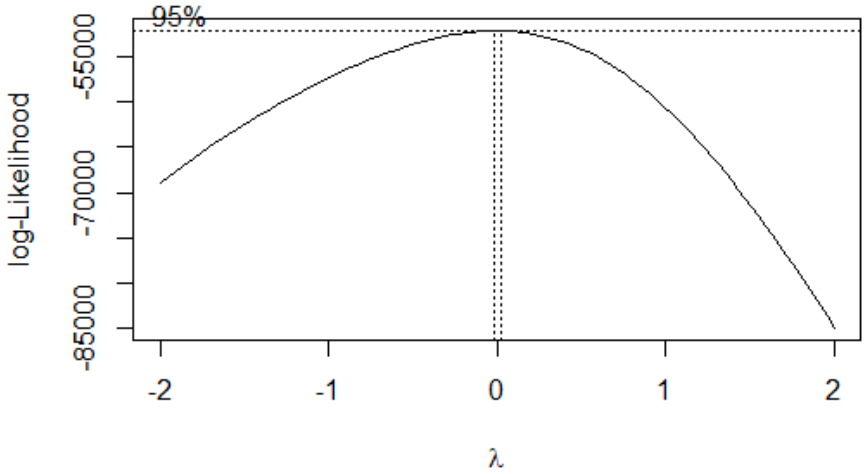
- Obvious triangular pattern
- Clear thick tails in Q-Q plot

=> Necessity of transformation on either predictors or response.



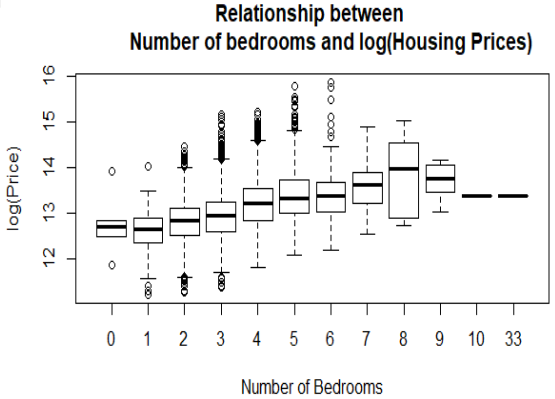


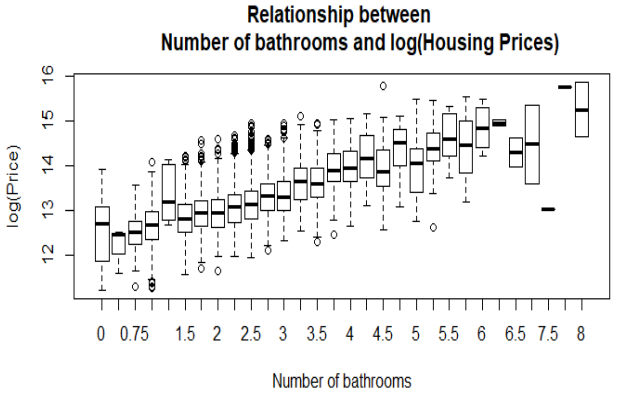
Transformation on Response

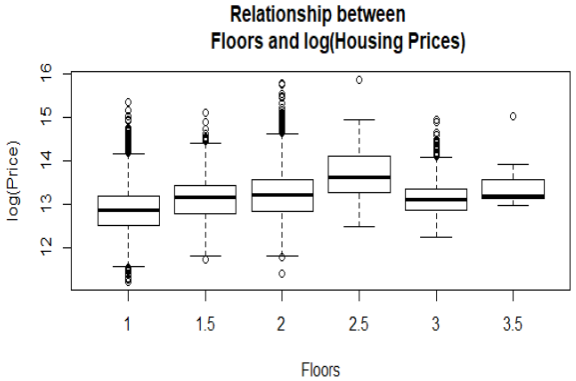


- Box-Cox method shows: $\lambda = 0$
- Logarithmic transformation on ‘Price’

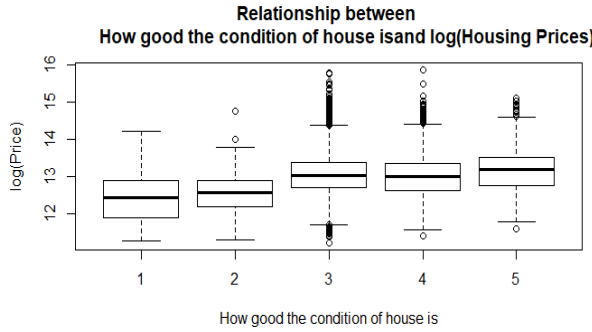
Transformation on Predictors

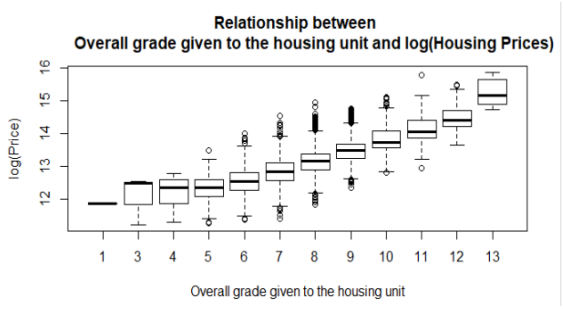


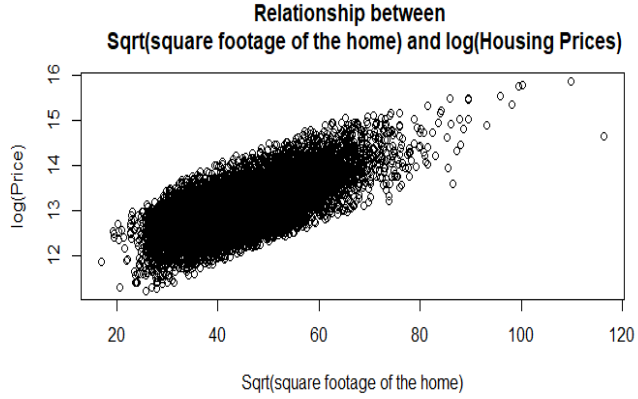


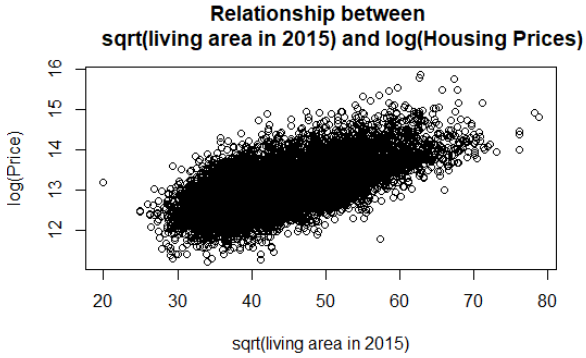


Transformation on Predictors









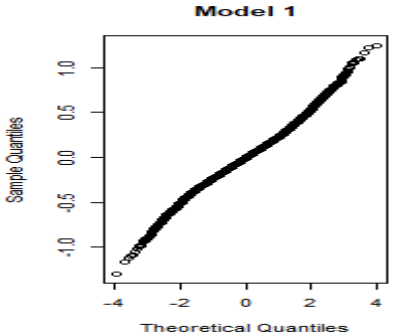
Model Selection

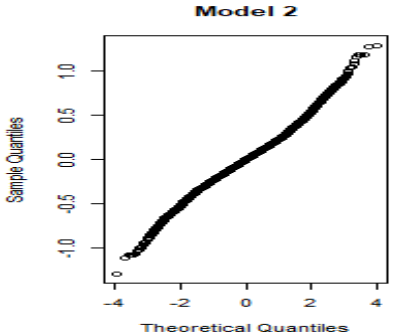
Method - Exhaustive approach

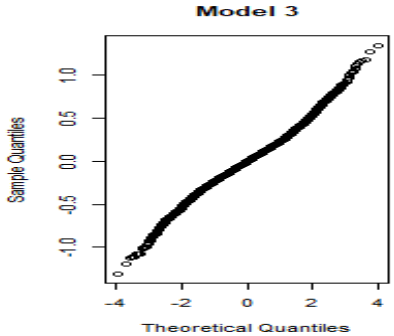
	Top 3 models on the basis of exhaustive method	SSRes	R2	AdjR2	MSE	Cp
Model 1	**log(Price)** ~ $\sqrt{Living\ area}$ + View + Condition + Grade + Year built + (Latitude - mean(Latitude))^2 + $\sqrt{Living\ area\ 2015}$	955.695	0.773	0.773	0.063	894.780
Model 2	**log(Price)** ~ $\sqrt{Living\ area}$ + View + Condition + Grade + Year built + (Latitude - mean(Latitude))^2	971.191	0.769	0.769	0.064	1152.309
Model 3	**log(Price)** ~ $\sqrt{Living\ area}$ + View + Grade + Year built + (Latitude - mean(Latitude))^2	990.385	0.765	0.765	0.066	1471.765

Model Selection

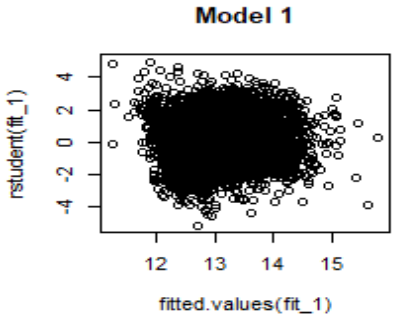
- **QQ-plots of top 3 models**

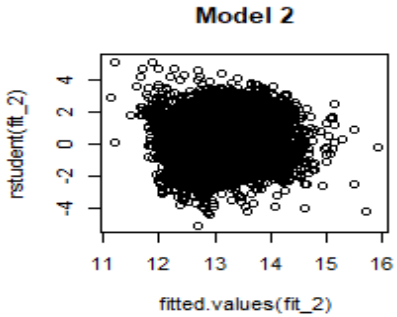


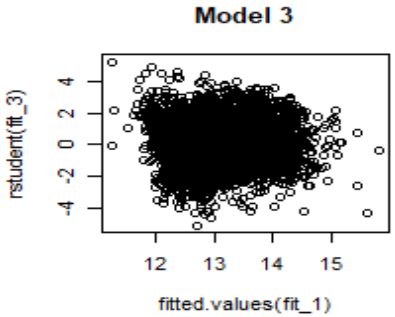




- **Residual plots**







Multicollinearity Analysis

view	condition	grade	yr_built	sqft_living_new
1.144334	1.167103	3.008840	1.569739	3.125492
lat_new.1	lat_new.2	sqft_living15_new		
1.086186	1.022219	2.613465		

Variance Inflation Factor:

- All the VIF values are less than 10.
- No correlation among predictors.

Final Model - Model 1

$$\log(\text{Price}) \sim \sqrt{\textit{Living area}} + \text{View} + \underline{\text{Condition}} + \text{Grade} + \text{Year built} + (\text{Latitude} - \text{mean}(\text{Latitude}))^2 + \sqrt{\textit{Living area 2015}}$$

- Cp = 894.7791(Lowest)
- MSE = 0.0632 (Lowest)
- R2adj = 0.7727 (Highest)
- QQ plot = Normally distributed
- Residual plot = Constant variance and no pattern

Final Model - R Output

```
Call:
lm(formula = log(price) ~ sqft_living_new + view + condition +
    grade + yr_built + lat_new + sqft_living15_new, data = train_hs)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29419 -0.15945 -0.00416  0.15042  1.24805

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.514e+01  1.685e-01   89.85   <2e-16 ***
sqft_living_new   1.809e-02  3.797e-04   47.65   <2e-16 ***
view              8.034e-02  2.836e-03   28.33   <2e-16 ***
condition         6.016e-02  3.397e-03   17.71   <2e-16 ***
grade             1.691e-01  3.010e-03   56.17   <2e-16 ***
yr_built         -2.398e-03  8.699e-05  -27.57   <2e-16 ***
lat_new1          2.353e+01  2.620e-01   89.79   <2e-16 ***
lat_new2         -8.211e+00  2.542e-01  -32.30   <2e-16 ***
sqft_living15_new 7.068e-03  4.514e-04   15.66   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2514 on 15120 degrees of freedom
Multiple R-squared:  0.7728,    Adjusted R-squared:  0.7727
F-statistic:  6430 on 8 and 15120 DF,  p-value: < 2.2e-16
```

Detecting Influential Observations:

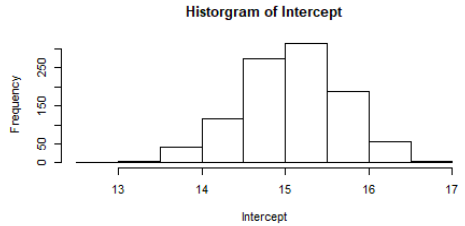
- Leverage points:
- Leverage cutoff value: 0.001189768 (k=8 n=15129 on training data set)
- Identified few leverage points

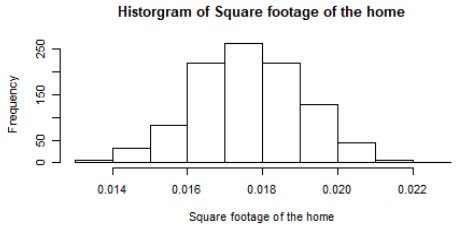
- Influential points:
- No Influential point present
- For all the points - Cook’s distance is less than 1

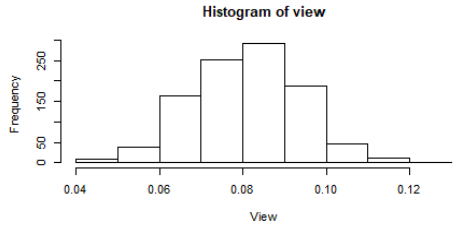
MSP on validation dataset

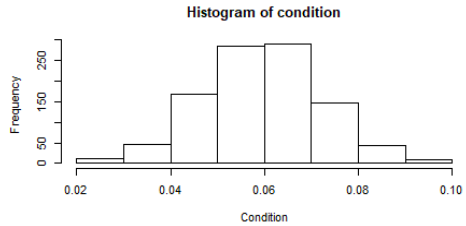
- Mean squared prediction error is 0.06122029.
- Mean squared residuals of our original model is 0.06
- Model is fairly successful at making predictions

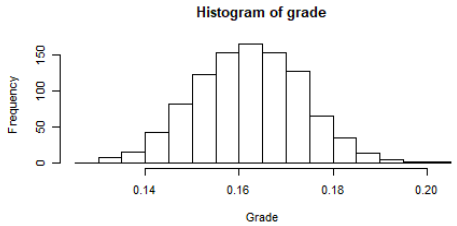
Bootstrap

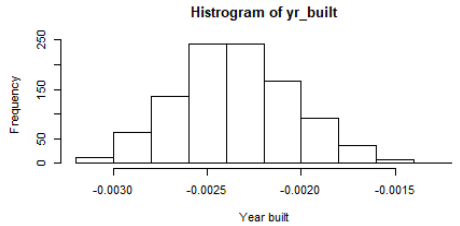

Historgram of Intercept
Frequency
0
50
150
250
13
14
15
16
17
Intercept

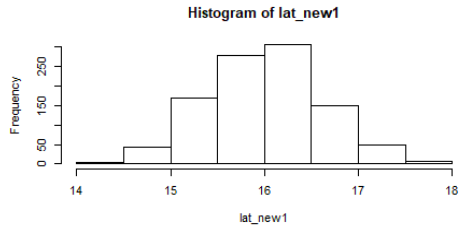

Historgram of Square footage of the home
Frequency
0
50
150
250
0.014
0.016
0.018
0.020
0.022
Square footage of the home


Histogram of view
Frequency
0
50
150
250
0.04
0.06
0.08
0.10
0.12
View

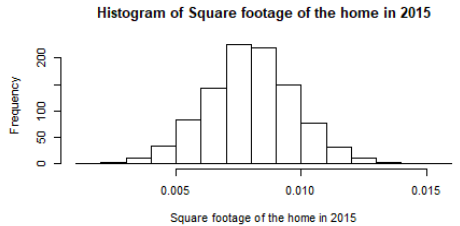

Histogram of condition
Frequency
0
50
150
250
0.02
0.04
0.06
0.08
0.10
Condition


Histogram of grade
Frequency
0
50
100
150
0.14
0.16
0.18
0.20
Grade

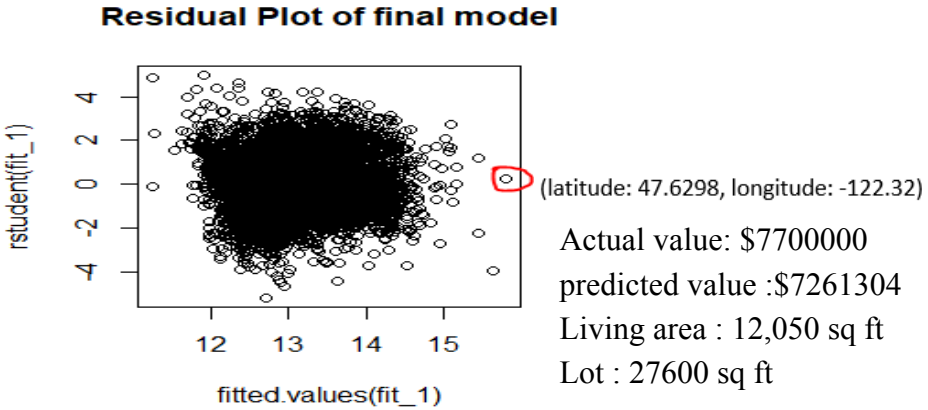

Histrogram of yr_built
Frequency
0
50
150
250
-0.0030
-0.0025
-0.0020
-0.0015
Year built


Histogram of lat_new1
Frequency
0
50
150
250
14
15
16
17
18
lat_new1

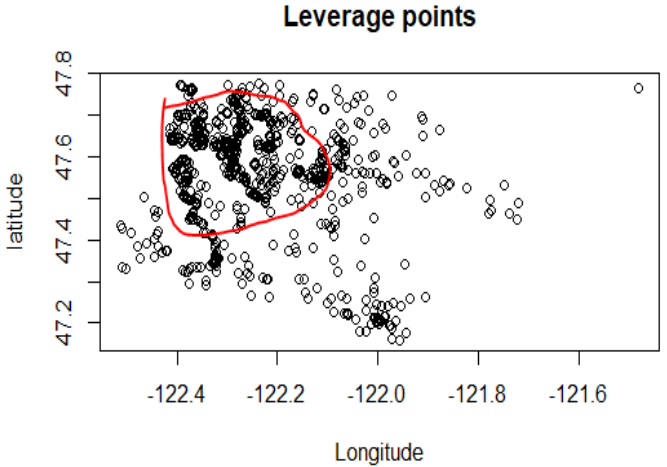

Histrogram of lat_new2
Frequency
0
50
150
250
-8
-7
-6
-5
-4
lat_new2


Histogram of Square footage of the home in 2015
Frequency
0
50
100
200
0.005
0.010
0.015
Square footage of the home in 2015

Exploration of Leverage Points



Actual value: $7700000
predicted value :$7261304
Living area : 12,050 sq ft
Lot : 27600 sq ft

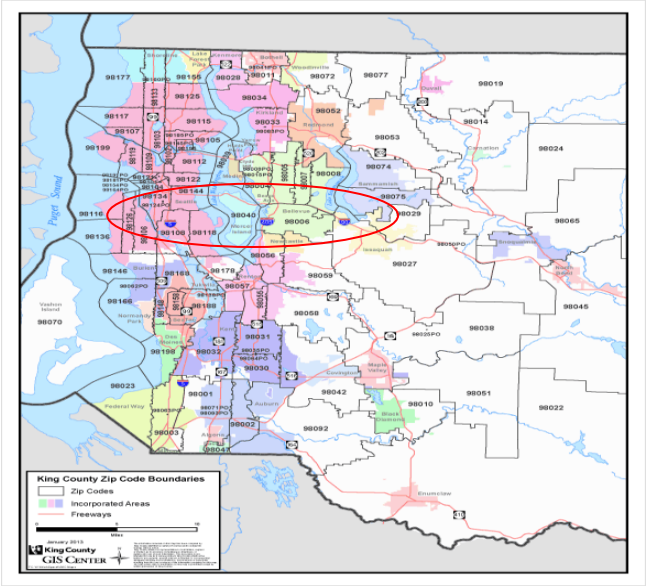


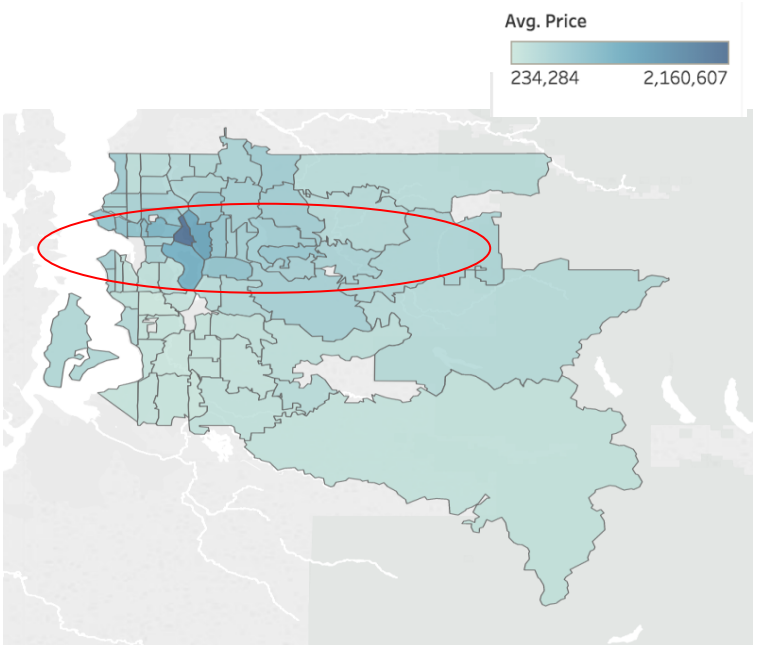
- Most of leverage points are falling in the latitude between (47.4 - 47.7)degree and longitude (-122.4, -122.1) degree.
- This zone has the highest price range.

Conclusion

- 77.27% of variability in housing prices can be explained through our final model
- As per model, 7 predictors have significant influence on price of house sold, which are **living area, view, condition, grade, year built, latitude** and **living area in 2015**.
- Unlike our initial guess, our final model does not include waterfront.
- Longitude of the house seems not to make an impact on the home values

Conclusion





- Prices are varying with latitude
- High priced houses are found near interstate highways

Future Study

- Take “Date” into consideration when we have a dataset with longer timespan.
- Take Environmental factors into consideration when they are provided in future study.

References

"Boundaries." *King County, 24 Apr. 2017,* *www.kingcounty.gov/services/gis/Maps/vmc/Boundaries.aspx*. Accessed 20 Nov. 2018.

Bremer, Martina. (2018). *Regression Theory and Methods*.

"Consumer expenditures 2017." *U.S. Bureau of Labor Statistics,* 13 Sep. 2018, www.bls.gov/news.release/cesan.nr0.htm. Accessed 20 Nov. 2018.

DePillis, Lydia. "How Washington could actually make housing more affordable." *Cable News Network,* 24 Jul. 2018, www.cnn.com/2018/07/24/politics/affordable-housing/index.html. Accessed 20 Nov. 2018.

"House Sales in King County, USA." *Kaggle, 2016,* www.kaggle.com/harlfoxem/housesalesprediction. Accessed 20 Nov. 2018.

"Kaggle." *Wikipedia,* 13 November 2018, en.wikipedia.org/wiki/Kaggle. Accessed 20 Nov. 2018.

"King County, Washington." *Wikipedia,* 20 November 2018, en.wikipedia.org/wiki/King_County,_Washington. Accessed 20 Nov. 2018.

Montgometry, Douglas, et al. *Introduction to Linear Regression Analysis.* 5th ed., John Wiley and Sons, Inc., 2012.

Thank you!

Any Questions?