
Cribs

Regression modeling for home price

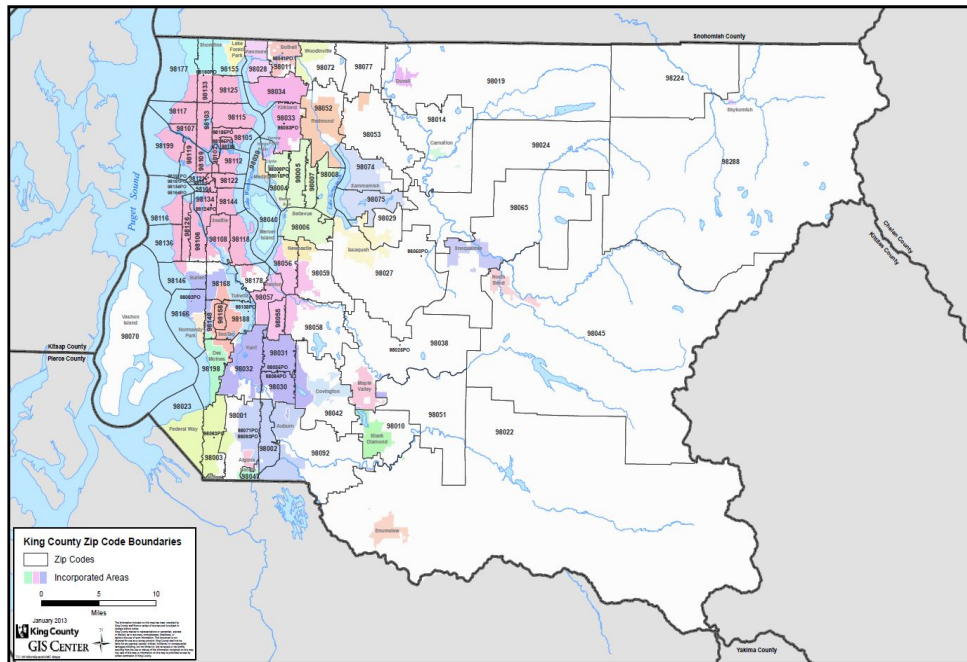
— Jun Tae Son, Omer Saif Cheema, —
Yusheng Zhu

Describe data and goal

- Downloaded from Kaggle.com
- 21,613 records of home sales in King County, WA. from May 2014 to May 2015.

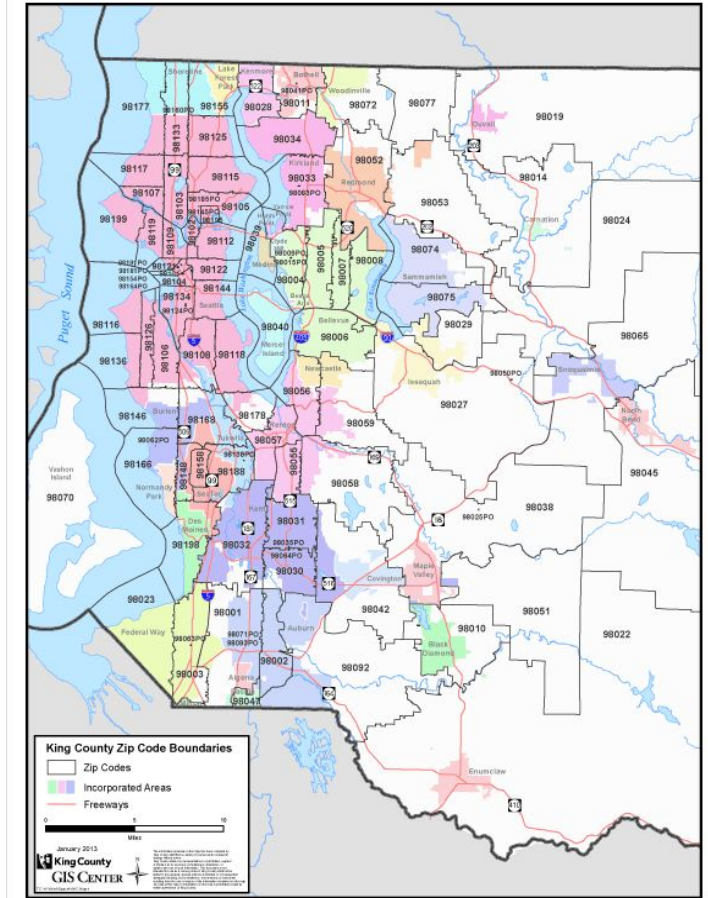
The goal of the analysis

- Generate the best fitted regression model for predicting home sale price in King county.



Houses in the West

- All the data was about houses located in the west side of King County.
- There were no houses in the data which were located at the east side of King County.



Features in the Dataset

- ID
- Date
- Price
- Bedrooms
- Bathrooms
- Sqft_living
- Sqft_lot
- Floors
- Waterfront
- View
- Condition
- Grade
- Sqft_above

- Sqft_basement
- yr_built
- yr_renovation
- zipcode
- lat
- long
- Sqft_living15
- Sqft_lot15

21 Variables

Features
used in
the
analysis



- Date
- Price
- Bedrooms
- Bathrooms
- Sqft_living
- Sqft_lot
- Floors
- Waterfront
- View
- Condition
- Grade
- Sqft_above

- Sqft_basement
- yr_renovation
- lat
- long

16 Variables

Explaining Variables

- Price = house price
- Bedrooms = # of bedrooms
- Bathrooms = # of bathrooms
- Sqft_living = square footage of the interior living space
- Sqft_lot = square footage of the land space
- Sqft_above = Sqft_living - Sqft_basement
- Floors = # of floors
- Waterfront = value is 1 if there is waterfront otherwise, 0 (binary)
- Condition = value is 1 if the condition is reasonable otherwise, 0 (binary)

We used the above variables as they were.

Explaining Variables

- `sqft_basement`: The square footage of the interior housing space that is below ground level

Created a dummy variable, `basement`, for `sqft_basement`.

`basement = 1` when `sqft_basement > 0` otherwise, `basement = 0`,

- `yr_renovation`: The year of the house's last renovation

Created a dummy variable, `renovated`, for `yr_renovation`

`renovated = 1` when `yr_renovation` is given otherwise, `renovated = 0`

Explaining Variables

Date: Date of the house sale

Original Format
YYYYMMDDT000000

Read in first
8 digits
YYYYMMDD

Assign each date to one of the four quarters in a year

Make Q1 the
base level

Create Q2, Q3 and Q4 as dummy variables

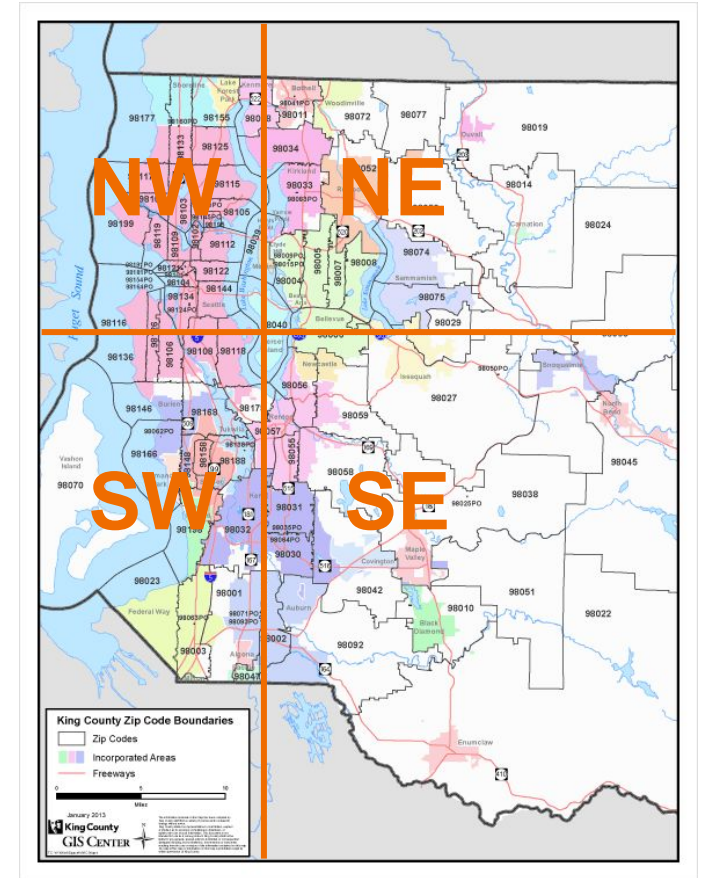
Location: Given in latitudes and longitudes

Find the median
longitude and latitude.

Assign houses to North West, South West, North East or South East.

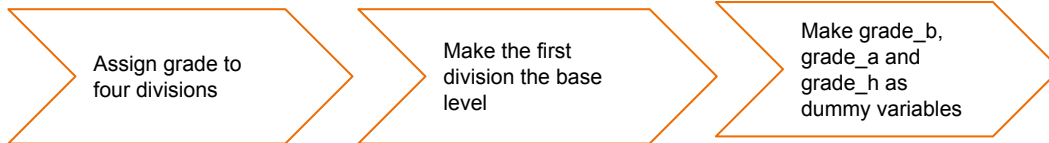
Make North West (NW) the base level

Create NE, SW, SE
as dummy variables



Explaining Variables

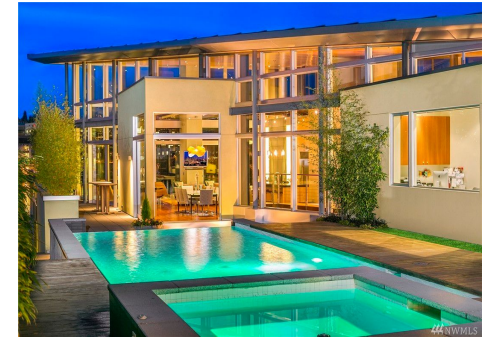
Grade: An index from 1 to 13, where 1-3 falls short of building construction and design, 7 has an average level of construction and design, and 11-13 have a high-quality level of construction and design



grade_b	grade_a	grade_h
grade = 4 - 6	grade = 7 - 10	grade = 11 - 13



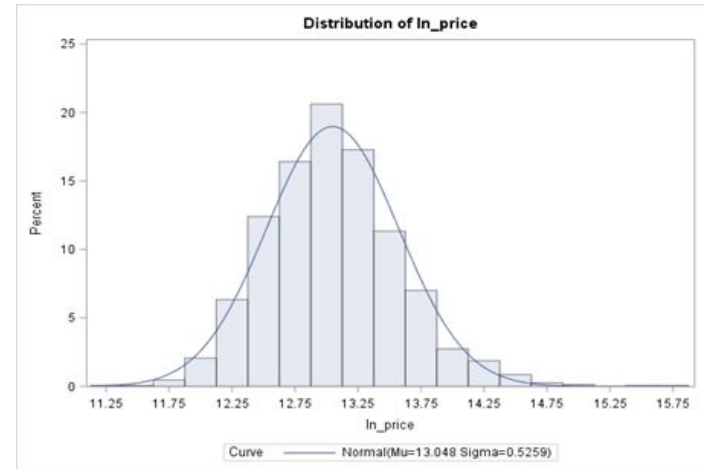
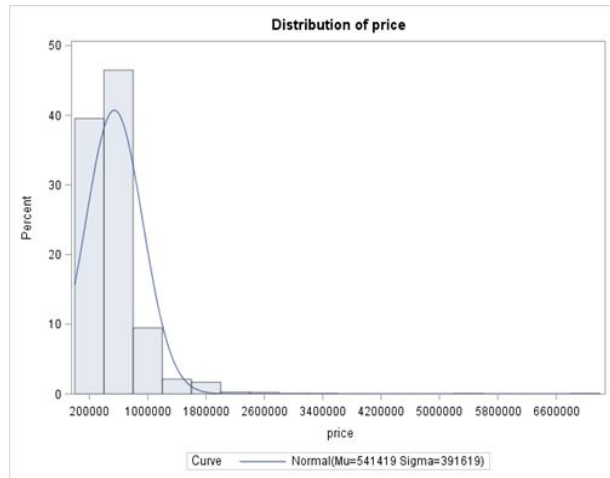
Grade = 1



Grade = 13

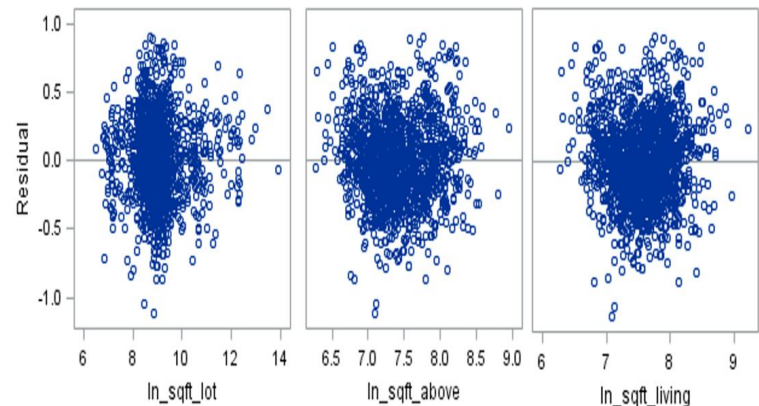
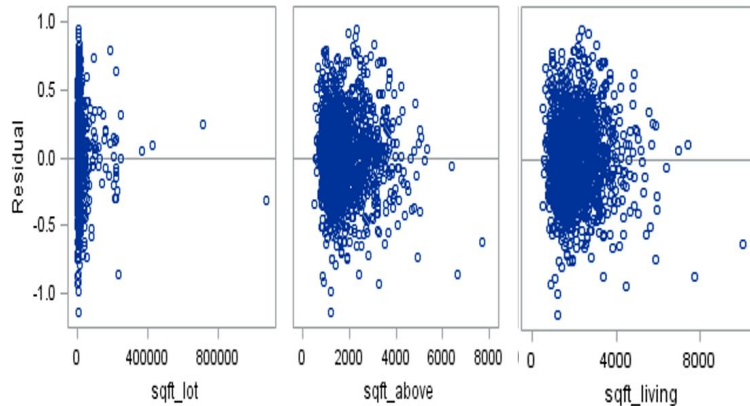
Transformation of Dependent Variable

We transformed the response variable, price, with log transformation because it was positively skewed and used \ln_price as our dependent variable.



Transformation of Independent variables

We transformed the `sqft_living`, `sqft_above` and `sqft_lot` because their residuals did not have constant variance and independence.



Methodology

- **STEP1:** Each member selected a **sample of 1500** rows randomly from the dataset.
- **STEP2:** Preprocessed the sample by creating dummy variables.
- **STEP3:** Transformed the response and three quantitative variables.
- **STEP4:** Checked the sample for **multicollinear** predictors and removed them.
- **STEP5:** Created **interactive variables**, however, they exhibited multicollinearity.
- **STEP6:** To fix the multicollinearity we **centered** the quantitative variables in the interactive terms and created new centered interactive variables.

Methodology

- STEP7: Afterwards, we identify **outliers and influence points** in our respective samples and take all or some of them out of the sample.
- STEP8: Then, we **split** the data into a training set and testing set.
- STEP9: Each member used two different **selection method** to find the best set of predictors using the training data set.



JUN'S PART



Methodology

- STEP10: Validity test on training set and **Performance test** on Testing set
- STEP11: Choosing the **final model**
- STEP12: Interpretation of the **important parameters**

After model selection

Model 1 (Jun): $\ln_price = 13.3336 - 0.0746 \cdot bedrooms - 0.6983 \cdot \ln_sqft_above_c + 0.7534 \cdot waterfront + 0.3226 \cdot view_good + 0.1536 \cdot condition_good - 0.1123 \cdot grade_b + 0.4532 \cdot grade_h + 0.3768 \cdot basement - 0.5468 \cdot SW - 0.3844 \cdot SE - 0.1054 \cdot Q4 - 0.2190 \cdot above_NE_c + 0.2811 \cdot above_SW_c - 0.2688 \cdot above_SE_c$

Model 2 (Omer): $\ln_price = 13.6272 - 0.6689 \cdot \ln_sqft_living_c + 0.3597 \cdot view_good + 0.15775 \cdot condition_good - 0.5265 \cdot grade_b - 0.4731 \cdot grade_a + 0.1918 \cdot renovated - 0.4127 \cdot SW - 0.3761 \cdot SE - 0.2520 \cdot \ln_sqft_living_SE_c$

Model 3 (Yusheng): $\ln_price = 13.16315 - 0.6982 \cdot \ln_sqft_living_c + 0.6044 \cdot waterfront + 0.4120 \cdot view_good - 0.4693 \cdot SW - 0.4133 \cdot SE - 0.0906 \cdot bb_c + 0.2633 \cdot bathliving_c + 0.1038 \cdot bedliving_c - 0.2716 \cdot living_SE_c$

Model Validity test on Training set

	<i>M1</i>	<i>M2</i>	<i>M3</i>
<i>The number of predictors</i>	14	9	9
<i>Goodness of Fit</i>	p<0.001	p<0.001	p<0.001
<i>RMSE</i>	0.23755	0.29616	0.30424
<i>R square</i>	0.7301	0.6762	0.6637
<i>Adjusted R square</i>	0.7263	0.6743	0.6610

Predictive Performance on Testing set

	<i>M1</i>	<i>M2</i>	<i>M3</i>
<i>RMSE</i>	0.2410	0.2993	0.2722
<i>MAE</i>	0.1962	0.2436	0.2209
<i>R square</i>	0.7665	0.6933	0.7179
<i>Adjusted R square</i>	0.7563	0.6857	0.7109
<i>Cross-validated R square</i>	0.0262 (<0.3)	0.0153 (<0.3)	0.0494 (<0.3)

Interpretation of Regression Coefficients

Parameter Estimates

<i>Variable</i>	DF	Parameter Estimate	Standard Error	t Value	Pr > t 	Standardized Estimate
<i>ln_sqft_above_c</i>	1	-0.69832	0.07163	-9.75	<.0001	-0.58981
<i>basement</i>	1	0.37681	0.03111	12.11	<.0001	0.37483

Effect of basement on Price

Parameter estimate =0.3768

$$100*(e^{0.3768} - 1) = 45.76\%$$

Assuming all other variables constant,
home sale price increases by 45.76%
if the property has a basement.



Effect of sqft_above on Price

sqft_above... too complicated predictor

1. Log transformation on Y and X
2. Interaction variable (NW NE SW SE)
3. Centered

Base = NW



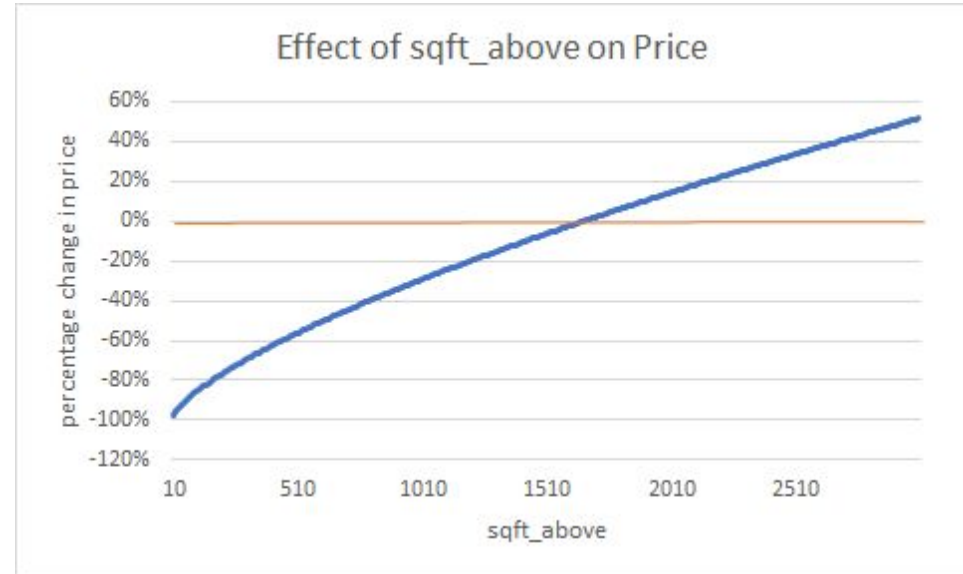
ln_PRICE

ln_sqft_above_c

Effect of sqft_above on Price

$$100 * (e^{(-0.6983 * (7.4107 - \ln(1)))} - 1) \\ = -99.43\%$$

$$100 * (e^{(-0.6983 * (7.4107 - \ln(1654)))} - 1) \\ = 0.01\%$$



Effect of sqft_above on Price

Home Sale Price will start to **increase**

IF the property is larger than 1,654 sqft

IF the property is located in NW

Yusheng's Part

Model Predictions



Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	<u>12.1495</u>	<u>12.5380</u>	0.0522	12.4353	12.6407	12.0624	13.0136	-0.3885
2	12.7038	13.1537	0.0454	13.0643	13.2430	12.6808	13.6266	-0.4499

ln_price: 0.3885



Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	<u>\$188,999.57</u>	<u>\$278,730.32</u>			\$173,234.25		\$448,471.31	
2	\$328,995.71	\$515,916.39			\$321,515.17		\$827,860.55	

House Price: \$89,730.75

Limitations

- Transformation Method may cause inaccurate prediction results
- Lack of domain knowledge
- Odd revelations on certain predictors' impact on house prices

Future Work

- Try different transformation method: Box-Cox
- Try on large dataset
- Hedonic Pricing Model--internal factors

--external factors



Summary



- 14 predictors
- Adjusted R square: 0.7563
- RMSE:0.2410
- Cross validated R square: $0.0262 < 0.3$
- Most important predictors: size of a house and whether there is a basement or not.
- If we had more variables like property taxes, neighbouring public schools, air quality, and interest rates we could come up with a more accurate model.



Q&A