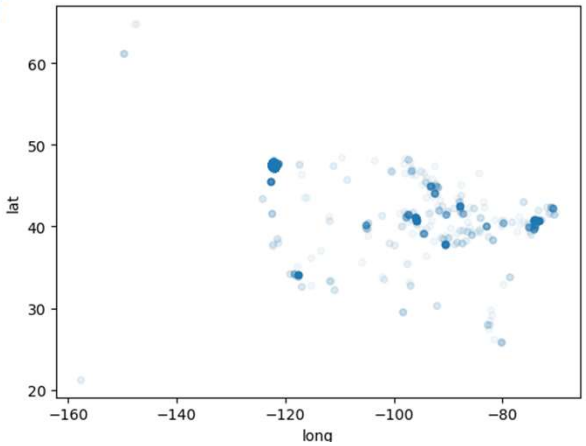
Analysis of housing prices

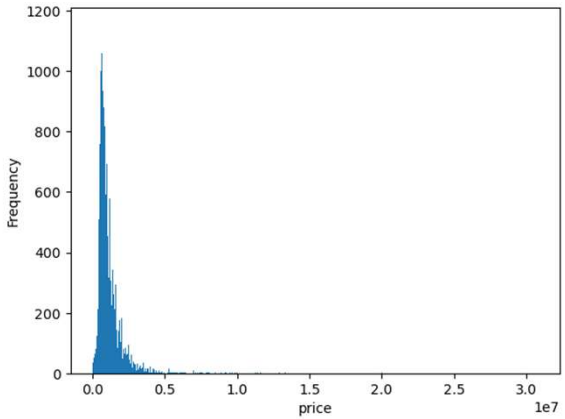
Business understanding

- The client is a housing planner
  - Must set prices and wants to use market data
  - It is necessary to know the impact on the housing price of various real estate metrics

Data understanding

- **Housing data from a Northwestern county and from the county government.**
- Key data categories include price, number of rooms, various square footage metrics, and age of the house.
- Each row of data represents a different house sold.
  - Within past few years
  - About 30,000 in data set.
- nearly all observations within Greater Seattle



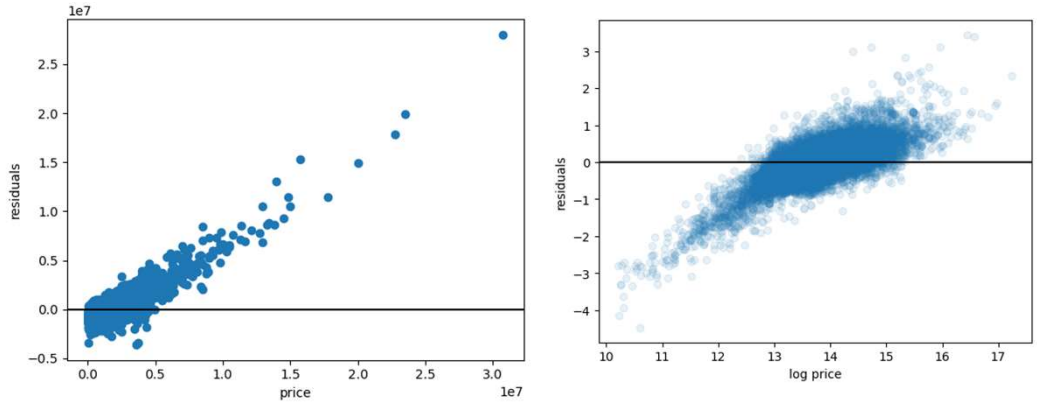


Modelling Overview

- Baseline model
  - 'sqft_living’
- Second model:
  - Add to previous, variables with price correlation greater than .25 and sqft_living correlation less than .75 to increase rsquared
- Third Model:
  - Log price/sqft living variables
  - Based on non-linearity issues (in residual plots and part regress) and non- normal issues in histograms
- Fourth Model:
  - More log tranformed variables (sqft_patio,sqft_garage)
  - to improve on linearity issue and heteroskedaticity issues and non-normality issues and to improve rsquared.
- Final model:
  - Log all numerical variables from prior model
  - Add categorical variable, Waterfront, to increase rsquared.
  - to improve on non-normality(despite improvement) and some heteroskedaticity

Results of iterative model process

- Final model heteroskedacities, linearity, and normality of residuals are improved from the baseline model.



- Residual plot appears normally distributed, and is improved from previous model, however J-B test still failed suggesting non-normality.

- Multicollinearity is low, all correlations below .75.

Final model

OLS Regression Results

Dep. Variable:	y_drop_X4	**R-squared:**	0.49	
		Adj. R-squared:	0.49	
		F-statistic:	2531	
		Prob (F-statistic):	0	
No. Observations:	15834			
	coef		**t**	**P>|t|**
const	6.7319		93.36	0
sqft_living_log	0.4971		37.352	0
sqft_garage_log	-0.1062		-10.208	0
sqft_patio_log	0.0423		10.332	0
WaterFront_Yes	0.3342		10.843	0
grade_num_log	1.7753		47.836	0
view_num_log	0.1138		12.842	0
Prob(Omnibus):	0	**Jarque-Bera (JB):**	32350.408	
Skew:	-0.969	**Prob(JB):**	0	
Kurtosis:	9.729	**Cond. No.**	249	

Results

- Model Evaluation
  - **Rsq is 0.49** compared to baseline of 0.38 and previous model of 0.46. This means the model accounts for **49%** of the variation in the dependent variable.
  - The mean squared error of the model is about **0.42.** This is a measure of how far off the predictions of log(price) are from the actual log(price).
- Interpretation of coefficients: All six predictor variables significant
  - WaterFront properties: **39.68% in price**
  - For each 1% increase in grade_num_log: **1.78% in price**
  - For each 1% increase in sqft_living_log: **0.50% in price**

Next Steps

- Try to improve on rsquared/reduce small non-linearity by using the lat/long scatter map to create dummy variables for specific geographic areas (would likely be about 10).
- Testing interaction variables (e.g. differing lot sizes and house sizes for different geographic areas.)

Thank you/Questions?