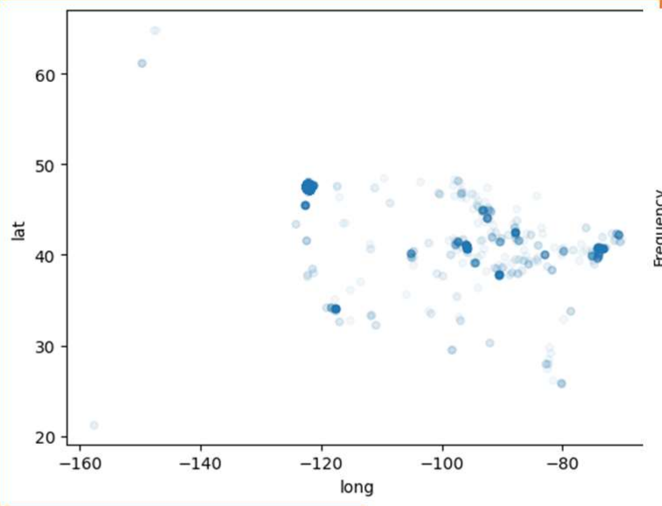
Analysis of housing prices

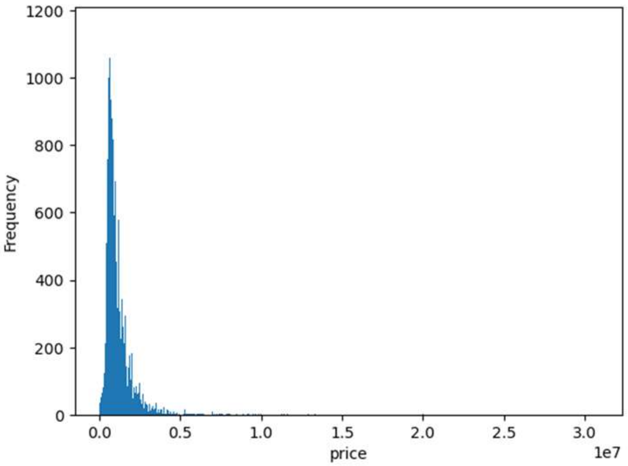
Business understanding

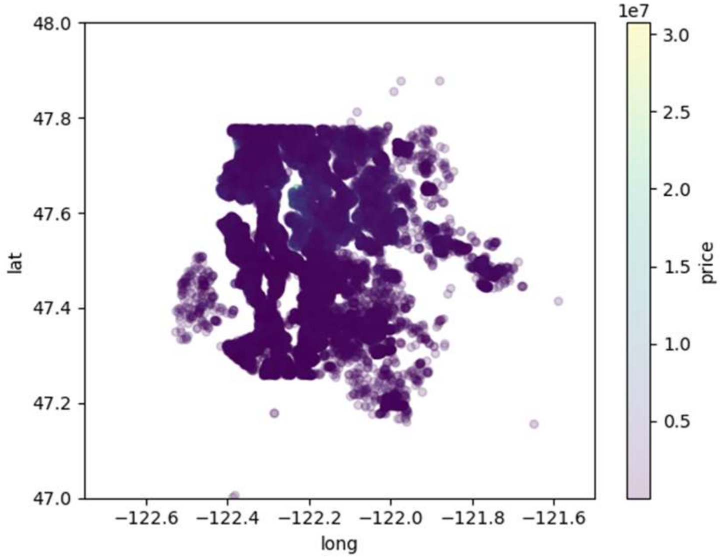
- The client is a housing planner
  - Must set prices and wants to use market data
  - It is necessary to know the impact on the housing price of various real estate metrics

Data understanding

- Housing data from a Northwestern county.
- Key variables: price, square footage and quality.
- 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, outliers cut.
  - high price center zone





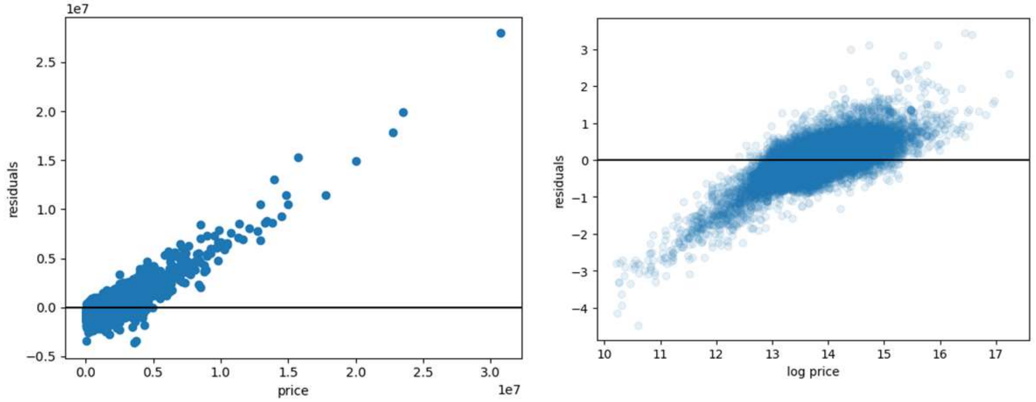


Modelling Overview

- 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 variables Waterfront and Jumbo 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:**	0.514
		Adj. R-squared:	0.514
		F-statistic:	2396
		Prob (F-statistic):	0
	coef	**P>|t|**	
const	**7.0719**	0	
sqft_living_log	**0.4832**	0	
sqft_garage_log	**-0.1035**	0	
sqft_patio_log	**0.0355**	0	
WaterFront_Yes	**0.3019**	0	
grade_num_log	**1.6657**	0	
view_num_log	**0.1033**	0	
Jumbo	**0.5701**	0	
Skew:	-1.084	Jarque-Bera (JB):	37618.359
Kurtosis:	10.233	Prob(JB):	0

Results

- Model Evaluation
  - **Rsq is 0.51** compared to baseline of 0.38 and previous model of 0.46. This means the model accounts for **51% of the variation** in the dependent variable.
  - The mean squared error of the model is about **0.41.** This is a measure of how far off the predictions of log(price) are from the actual log(price).

- Interpretation of coefficients: All seven predictor variables significant
  - Jumbo area: **76.84% in price**
  - 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

- Establish a better interpretation of the mean squared error.
- Further analyze the negative coefficient of garage size variable.
- Testing interaction variables (e.g. differing lot sizes and house sizes for different geographic areas.)

Thank you/Questions?