

The background of the slide is an abstract digital artwork. It features a series of overlapping, wavy lines in shades of red, orange, yellow, and green. These lines create a sense of depth and movement, resembling a stylized landscape or a complex data visualization. A bright, circular light source is positioned in the upper center, casting a soft glow across the scene and creating a lens flare effect. The overall color palette is warm and vibrant, with a gradient from light yellow at the top to deep red and purple at the bottom.

Analysis of housing prices

A large orange shape on the left side of the slide, consisting of a rectangle with a quarter-circle cutout at the top right corner.

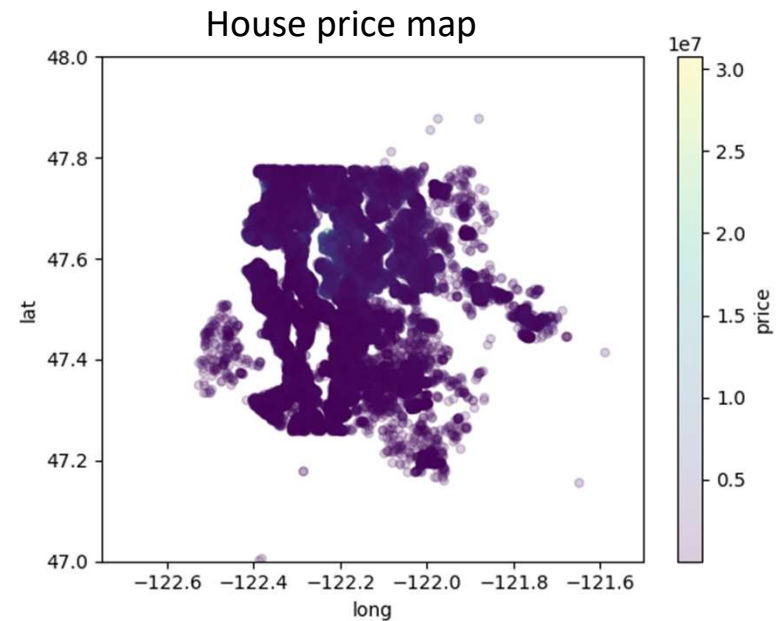
Business understanding

- 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

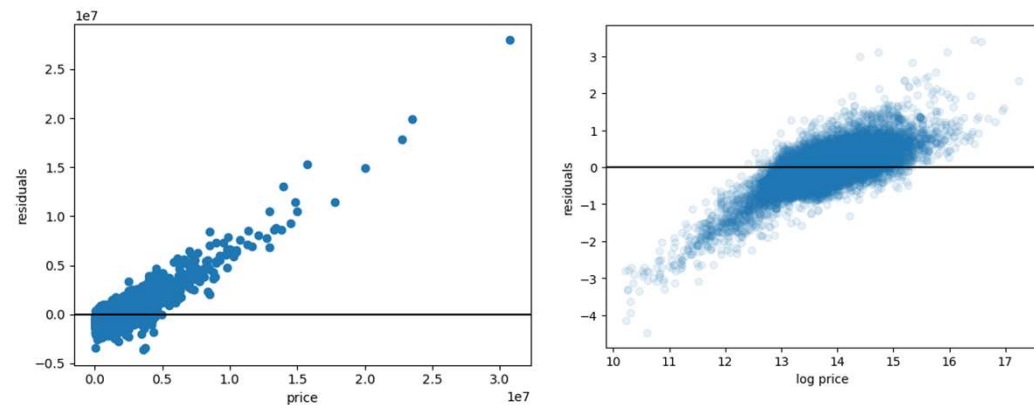


Modeling Overview: Linear regressions

- Baseline model
 - 'sqft_living'
- Intermediate models:
 - Add new numeric variables and conservatively log transform move variables with each iteration.
 - To improve on linearity issue and heteroskedasticity 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 heteroskedasticity

Final model
regression model
assumptions

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



- Residual histogram 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 & Results

OLS Regression Results

	coef
const	7.07
sqft_living_log	0.48
sqft_garage_log	-0.10
sqft_patio_log	0.04
WaterFront_Yes	0.30
grade_num_log	1.67
view_num_log	0.10
Jumbo	0.57

Results contd.

- Model Evaluation

- Rsq: **0.51**

- This means the model accounts for **51% of the variation** in the dependent variable.
 - Compared to baseline model of 0.38.

- Mean squared error: **0.41**.

- This is a measure of how far off the predictions of $\log(\text{price})$ are from the actual $\log(\text{price})$.

- Root mean squared error: **0.64**

- This is about the average of how far off the predictions of $\log(\text{price})$ are from the actual $\log(\text{price})$.

Interpretation of Coefficients

<u>Interpretation of coefficients table</u>			
Categorical Variables		% Effect of its presence on price	
WaterFront		35.24	
Jumbo Area		76.84	
Numeric Variables		% Effect of its 1% increase on price	
sqft_living		0.48	
sqft_garage		0.10	
sqft_patio		0.04	
grade_num		1.67	
view_num		0.10	

Recommendations

- To determine a price of a house, take a similar house with about 10% less sqft of living area and add 4.71% to that price.
- To determine a price of a house with a waterfront, take a similar house without a waterfront and add 35.24% to the non-waterfront price.
- To determine a price of a house in the Jumbo area, take a similar house not in the Jumbo area and add 76.84% to the non-Jumbo price.

Next Steps

- Establish a better interpretation of the root 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?