Multiple Linear Regression

Model-informed
recommendations for real estate

Why Linear Regression?

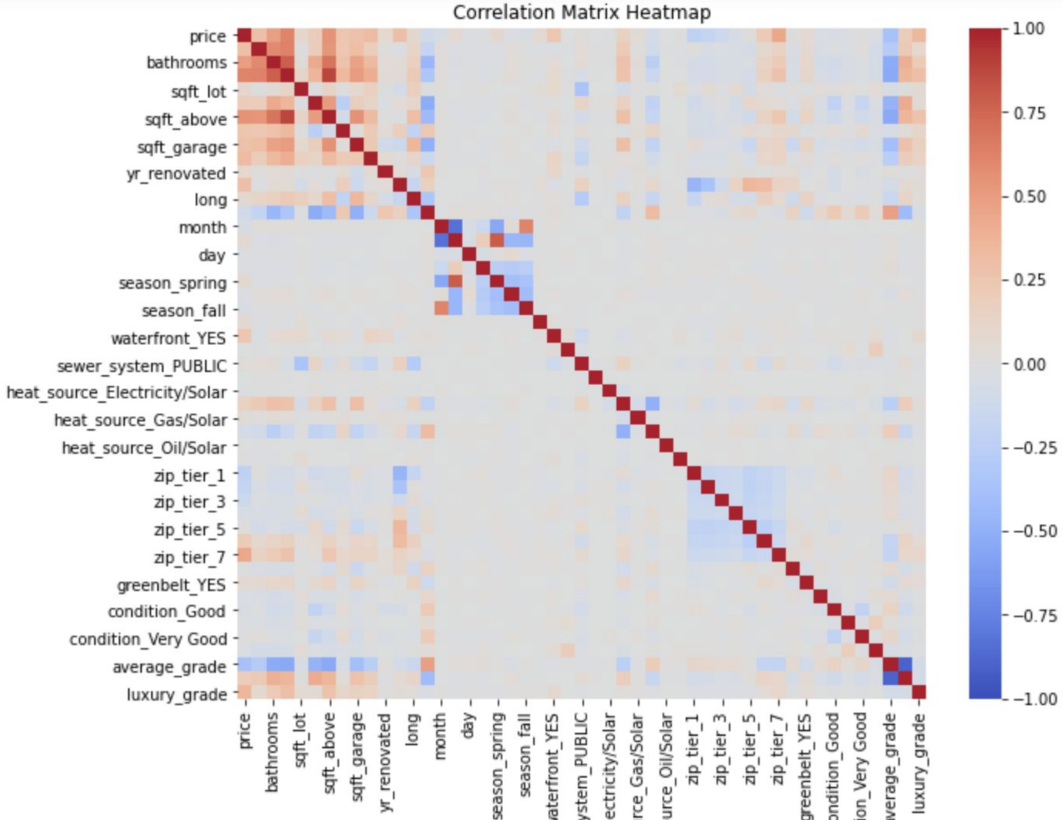
- Statistical analyses allow us to formally measure relationships

- Regression analysis estimates importance of each variable in predicting our target (price)

- We want to see which predictors have the strongest relationship to the sale price of houses in our new terf, King County.

Objective: get the lay of the land.

Clean data | Engineer features | Baseline model | Transform and scale data | Add/subtract features | Iterate model | Final model



Correlation heatmap to help us choose features

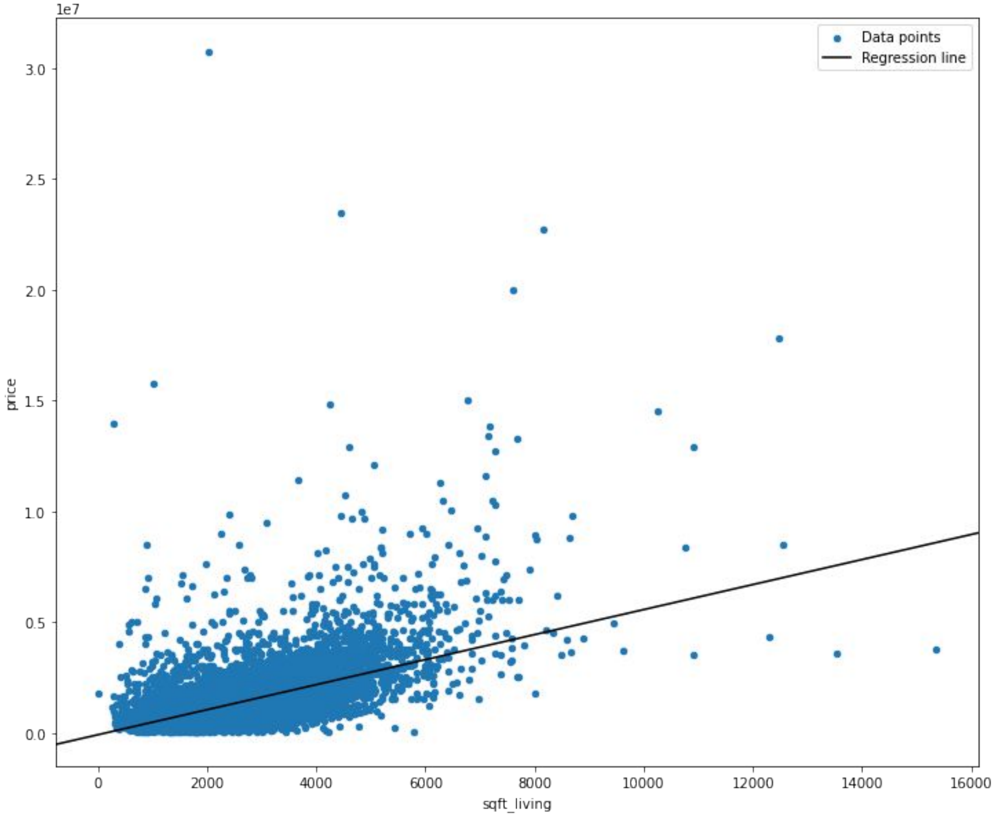
Feature Engineering

- Price per square foot
- House age
- Seasons
  - Winter, spring, summer, fall
- Consolidation of ‘grade features
  - Luxury, good, average, poor
-

Baseline Model

$y=mx+b$

Our first model will be a simple linear regression, which will use a selected, highly correlated feature from the dataset to measure against the raw data of our target, which is ‘price’.



Plot of raw ‘sqft_living’ against target ‘price’

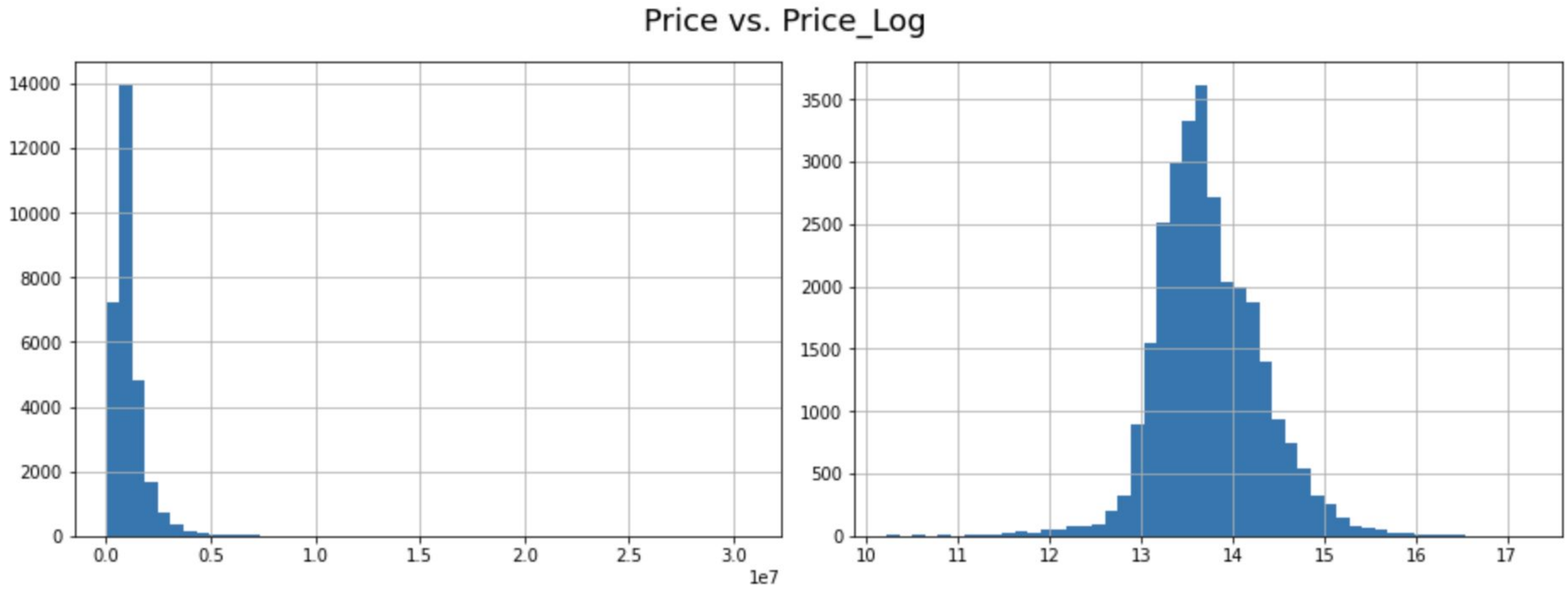
Multiple Linear Regression: All-relevant-features model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \ldots + \beta_i X_i$$

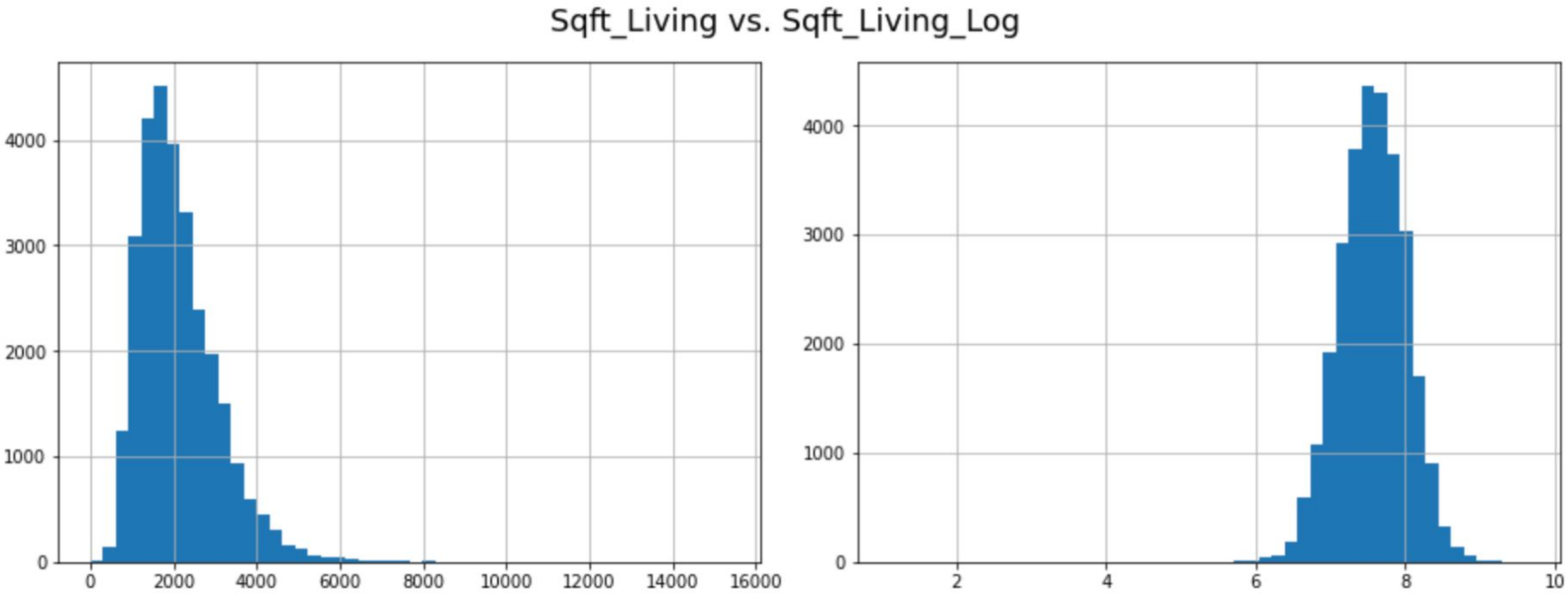
Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

Our second model will use the results of the baseline model to inform whether or not we want to:

- Transform the feature or the target feature we used
- Scale the data
- Add any or all relevant features to the model



Raw ‘price’ vs ‘price_log’ distribution



Raw ‘sqft_living’ vs ‘sqft_living_log’ distribution

MLR Model Results

- The model included all relevant features, including both features we engineered and one-hot encoded original features
    - zip_tier(s) 2-7
    - season(s)_winter, spring, summer, fall
    - house_age
- The features included in modeling explain about 70% of the variation in price
- Some features included are highly correlated with other features in the model, indicating undesirable multicollinearity

Features we will drop:

- season_winter
- sewer_system_PUBLIC RESTRICTED,
- heat_source_Oil
- heat_source_Oil/Solar
- heat_source_Other
- greenbelt_YES
- Luxury_grade,
- Sqft_garage_log
- ‘Lat’
- ‘Long’
- ‘Bedrooms’ (anomalous feature)

Final model: iterated MLR

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ \dots + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

Our final model will drop all aforementioned features to avoid multicollinearity, as well as anomalous ‘bedrooms’ feature that for some reason had a negative regression coefficient.

Final MLR Model Results

Our final model:

- Does not detect multicollinearity

- Explains 70% of the variation in price

- Performs moderately well, according to error analysis, but could perform better with the addition of interaction terms

- Identifies several ‘sqft’ and ‘zip_tier’ features as the strongest predictors of price variability.

Location (zip code tiers) is key.

- pp_sqft: 0.632
- sqft_above: 0.302
- sqft_living_log: 0.216
- zip_tier_7: 0.125
- zip_tier_6: 0.092
- sqft_basement: 0.077
- zip_tier_5: 0.075

Interaction plot between 'sqft_living_log' and top 3 zipcode tiers

Recommendations

Focusing on the top 3 zipcode tiers, final model coefficients indicate:

- A zip tier 5 property would sell for an average price 7.9% higher than a property in zip tier 4.
- A zip tier 6 property would sell for an average price 9.64% higher than a property in zip tier 5.
- A zip tier 7 property would sell for an average price 13.3% higher than a property in tip tier 6.

For appraisals:

1. Calculate the mean price of homes sold in that zipcode
2. From this, determine the zipcode tier of the home
3. Determine the expected percentage increase in mean price compared to the zip tier(s) below

Thank you! Any questions?

Let’s connect!

github.com/madelinebirch
mbirchhn@gmail.com
linkedin.com/in/madeline-birch-164000b5/