# 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.

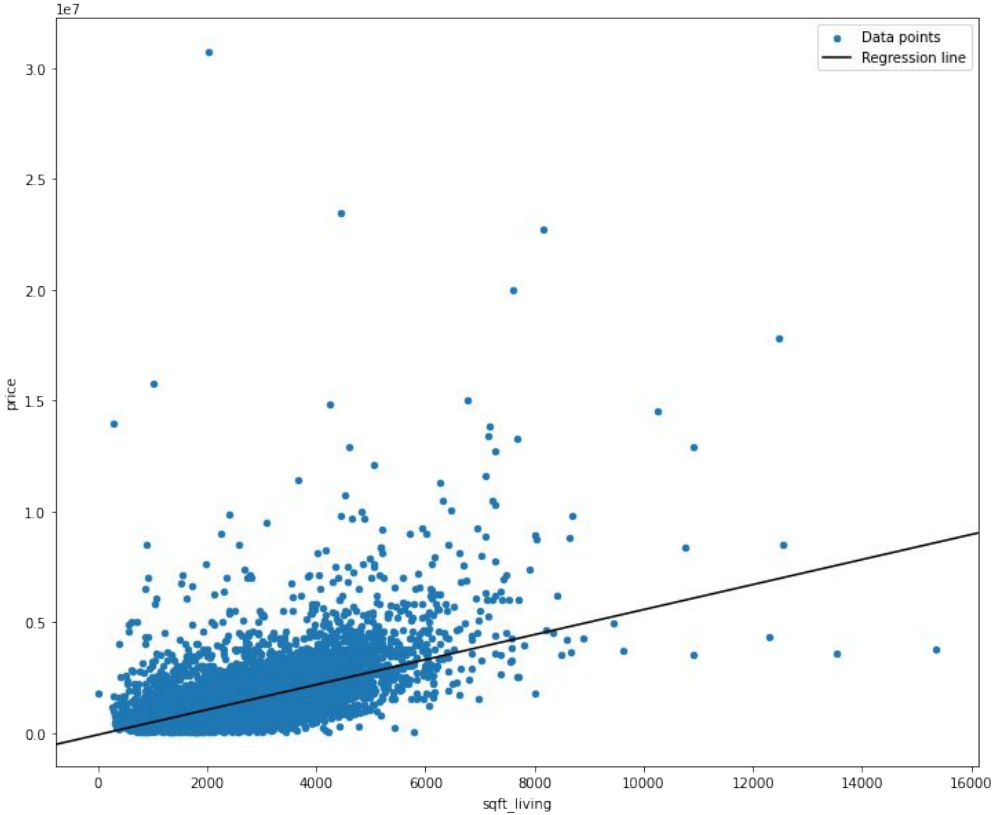
# Objective: get the lay of the land.

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

# 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

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

Y : Dependent variable
$\beta_0$ : Intercept
$\beta_i$ : 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 and all relevant features to the model

# 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

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

Y : Dependent variable
$\beta_0$ : Intercept
$\beta_i$ : 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
- Is well-fitted, according to error analysis
- Identifies several ‘sqft’ and ‘zip_tier’ features as the strongest predictors of price variability:

```
Predicted percentage increase in price_log for pp_sqft: 88.14%
Predicted percentage percentage increase in price_log for sqft_above: 35.26%
Predicted percentage percentage increase in price_log for sqft_living_log: 24.11%
Predicted percentage increase in price_log for zip_tier_7: 13.31%
Predicted percentage increase in price_log for zip_tier_6: 9.64%
Predicted percentage increase in price_log for sqft_basement: 8.00%
Predicted percentage increase in price_log for zip_tier_5: 7.79%
```


Interaction plot between ‘sqft_living_log’ and top 3 zipcode tiers

Next Steps:
Focus on location (zip code tiers) and square footage.