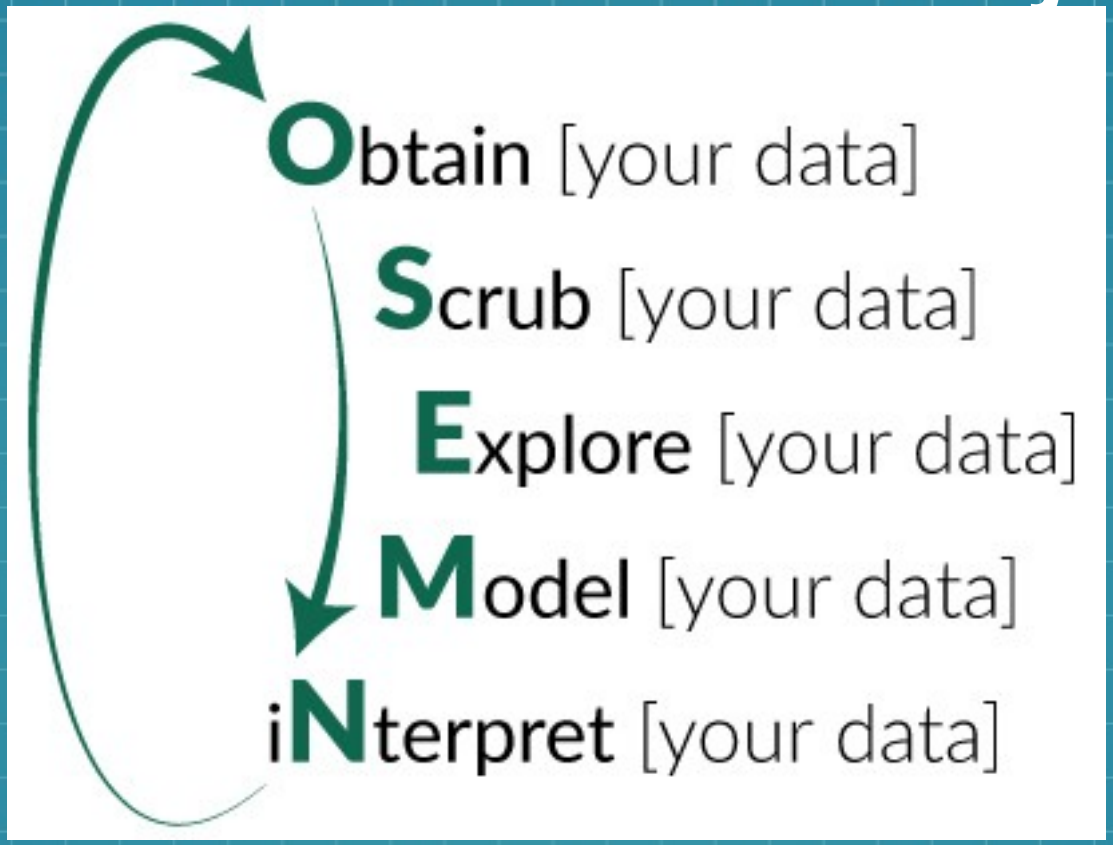
A Data Exploration of King County, WA

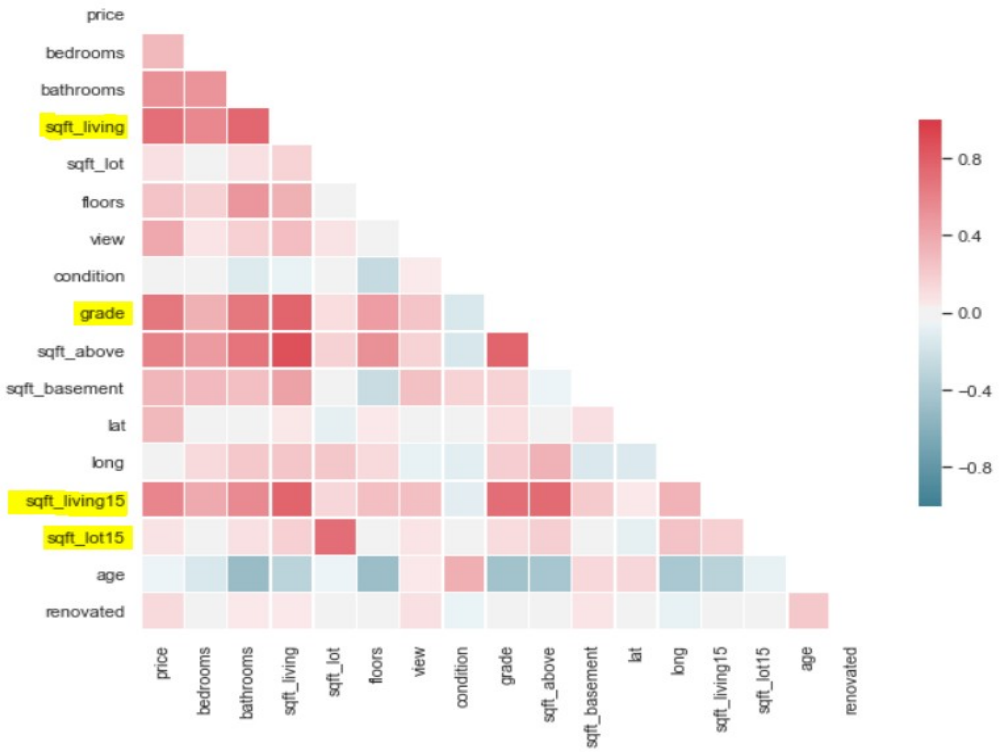
Framework of Project



Tweaks:

- a Re-scrubbing stage
- Experimentation with different Scrubbing and Modeling techniques to find the best outcome to proceed with
- This presentation *is* the **Interpret**ation stage

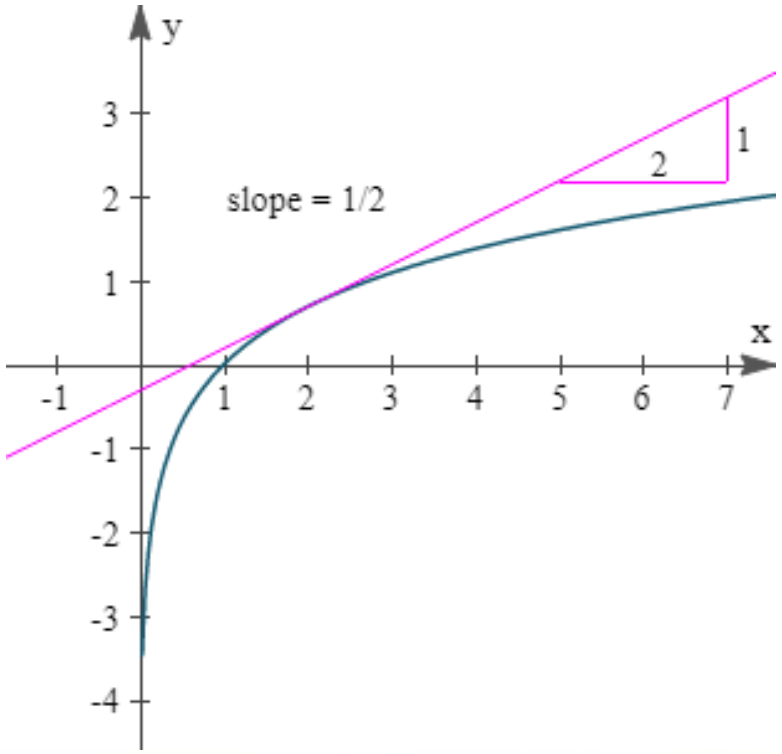
Methodology: Multicollinearity



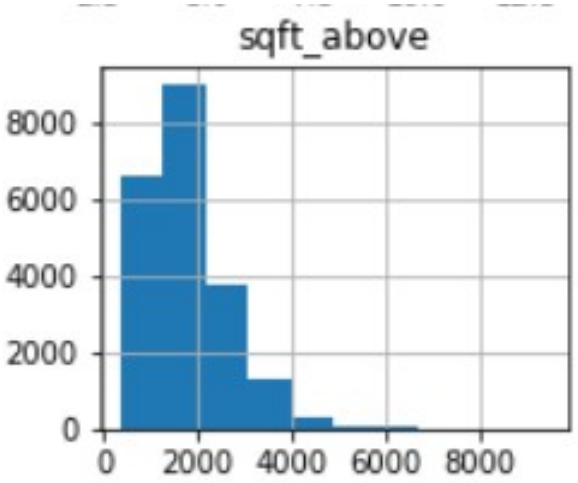
- Features already correlated with other features: sqft_living, grade, sqft_living15, sqft_lot
- Once removed, the model is more accurate

Use of Logarithms

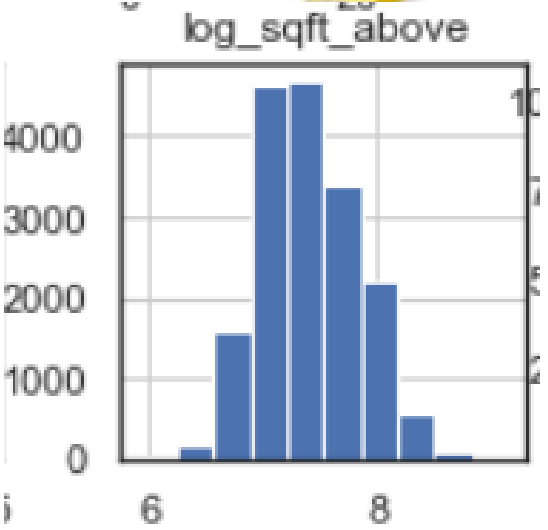
- Used on the Target and some predictors
- Expedites the statistical requirements of the algorithms used (Linear, Normal, Homoscedastic)
- Can give you information on general trend
  - But not specific Prices (e.g. “$200,000 is the best house”)



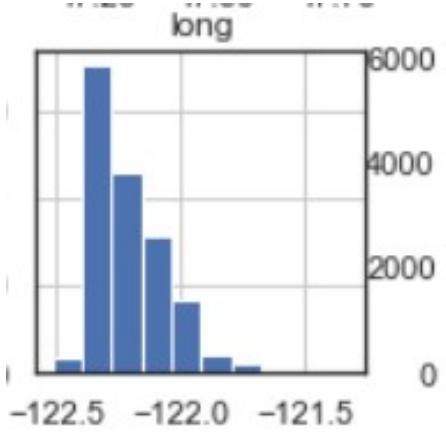
Methodology: Feature Scaling

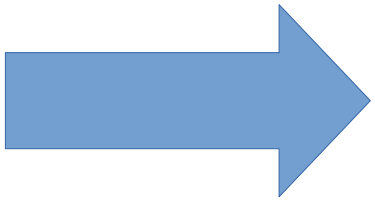


Log_{10}

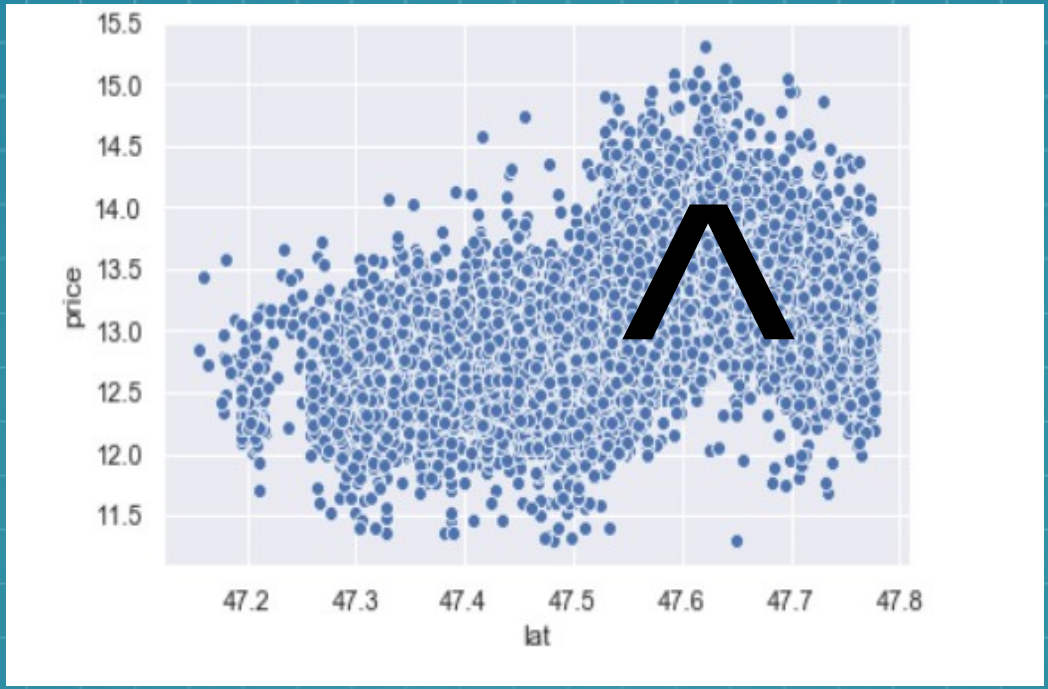


- Used to normalize any features with skewed distributions





on Latitude



- caret shape, unique to Latitude variable
- 47.62°N, 122.3321°W = Cascade Neighborhood of Seattle
  - Buy and improve any cheap homes here

Further Exploration

- Why does *lat* (and not *long*) have caret shape
- More insights by splitting data into Price brackets

$ ↔ $$ ↔ $$$

- Tweak features for better r^2