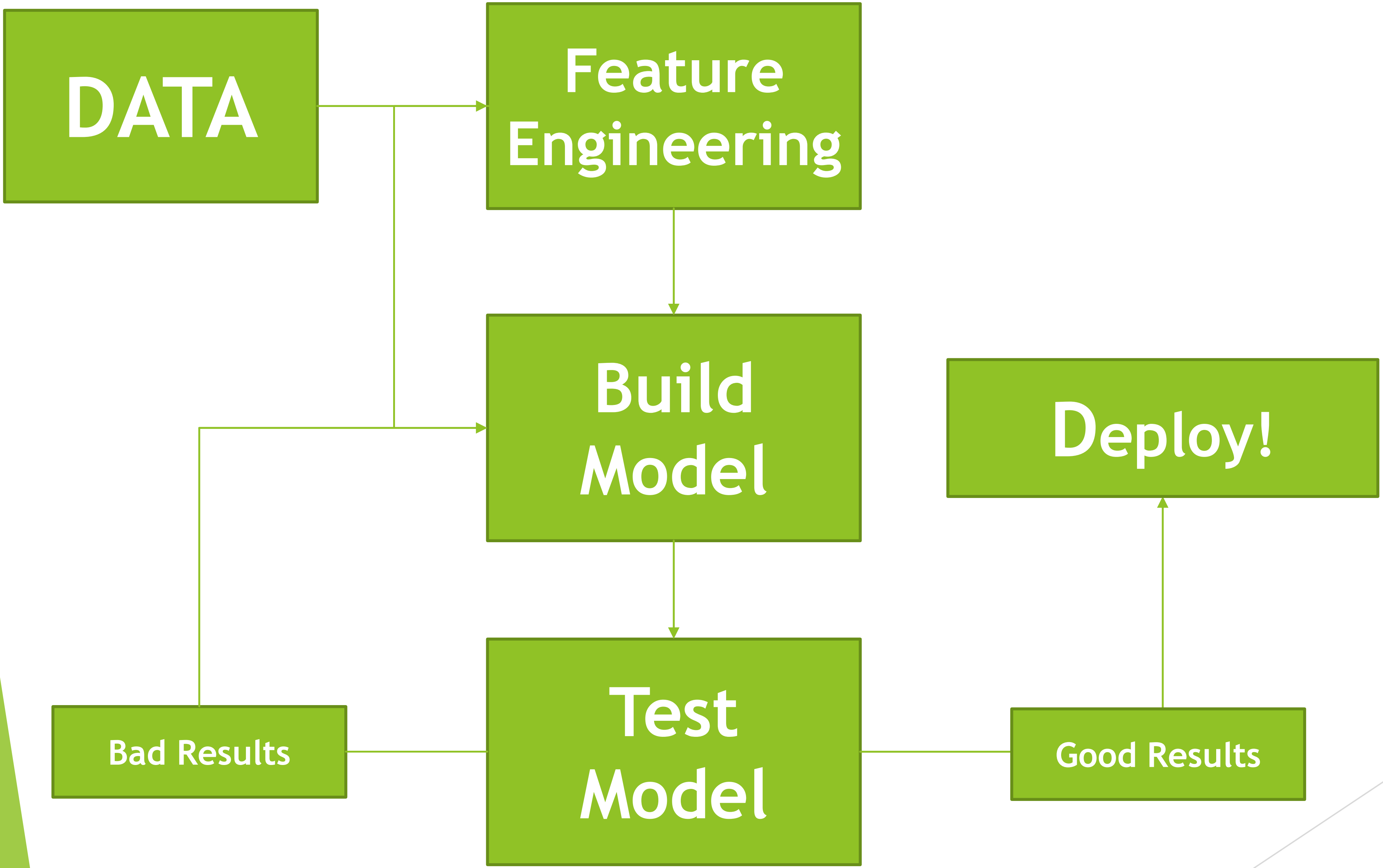
Linear Regression Modeling

King’s County Housing

The Data

- Target:
  - Goal is to predict: **House Price**

- Predictor Variables:
  - Number of Bedrooms, Bathrooms, and Floors
  - Sq. footage – Internal and Lot
  - Overall Condition (1-5)
  - Geographical Location (Based on Clustering)
  - Waterfront (yes / no)
  - Year Built
  - Renovated (yes / no)



DATA

Feature Engineering

Build Model

Test Model

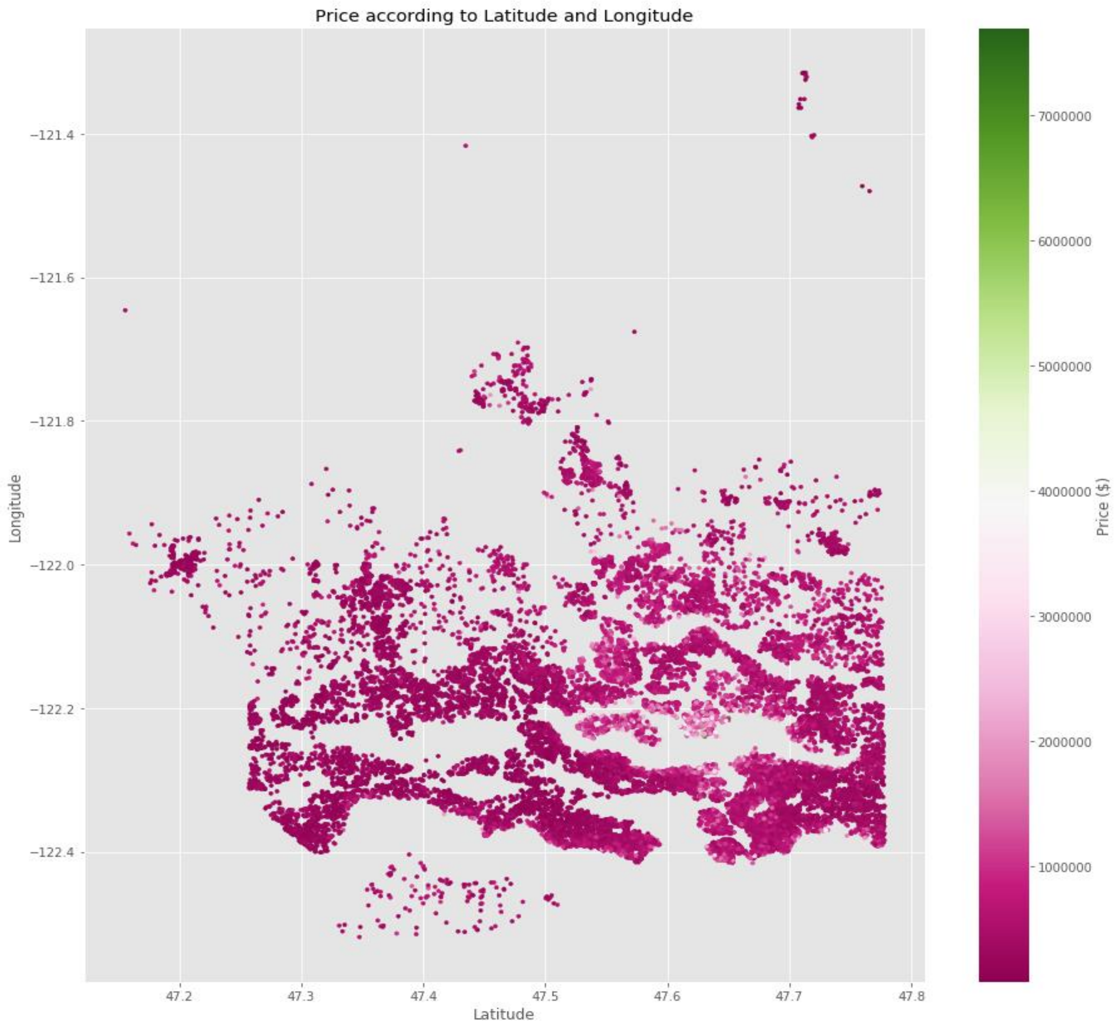
Bad Results

Good Results

Deploy!

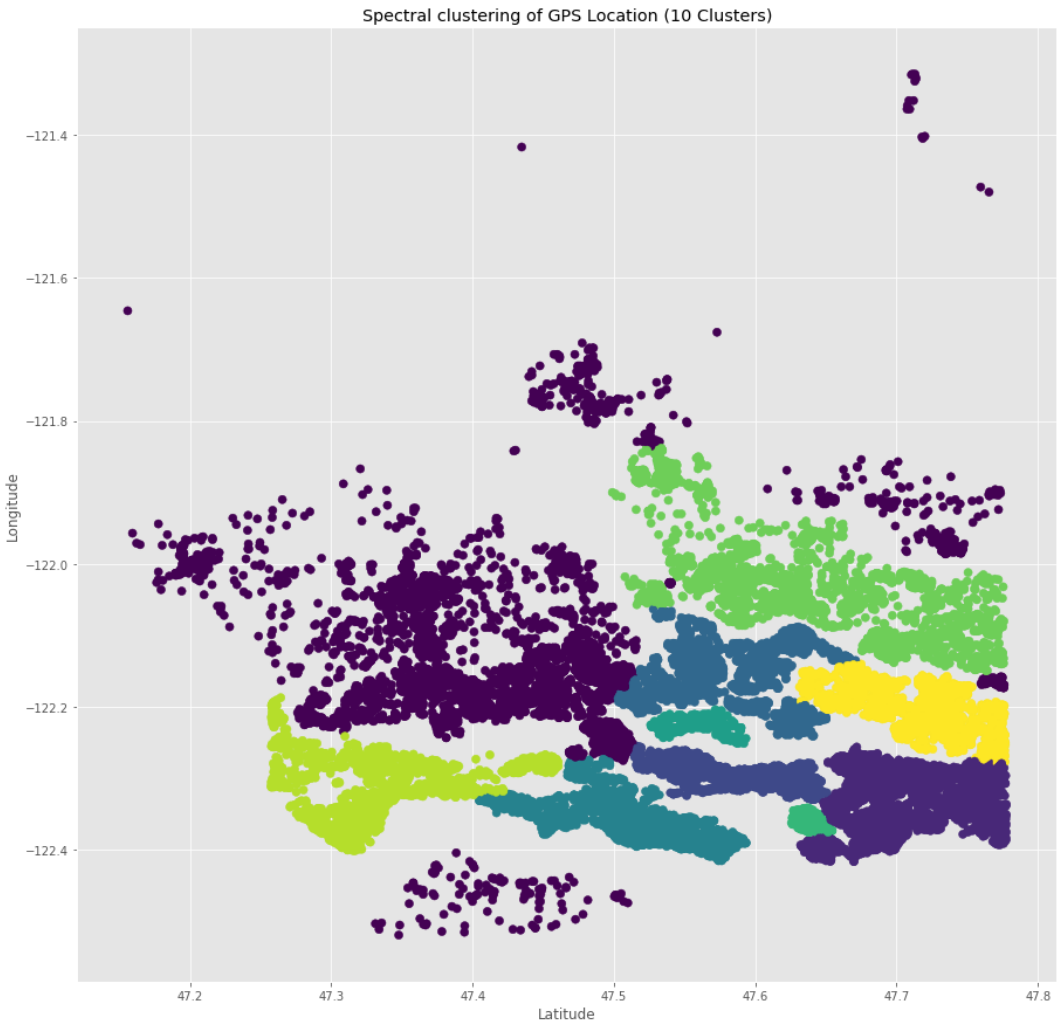
Problem: Dealing with Location Data

- Price depends on exact location
- Difficult to use latitude, longitude, zip code in linear model
- Still want to capture affect of location in model



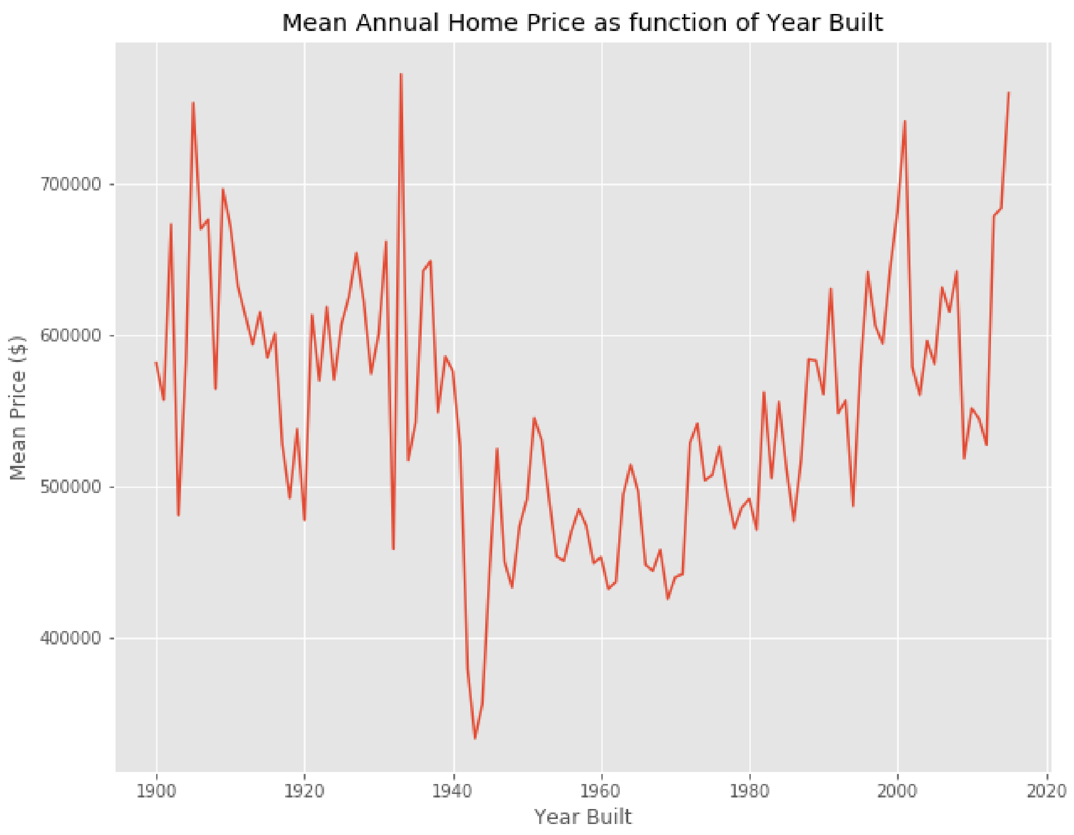
Solution: Spectral Clustering

- Finds patterns in data and separates into groups
  - Works well with complicated patterns
  - Gives good way to distinguish geographical affect of price
- Works well in linear model
  - Extracts meaning from location data



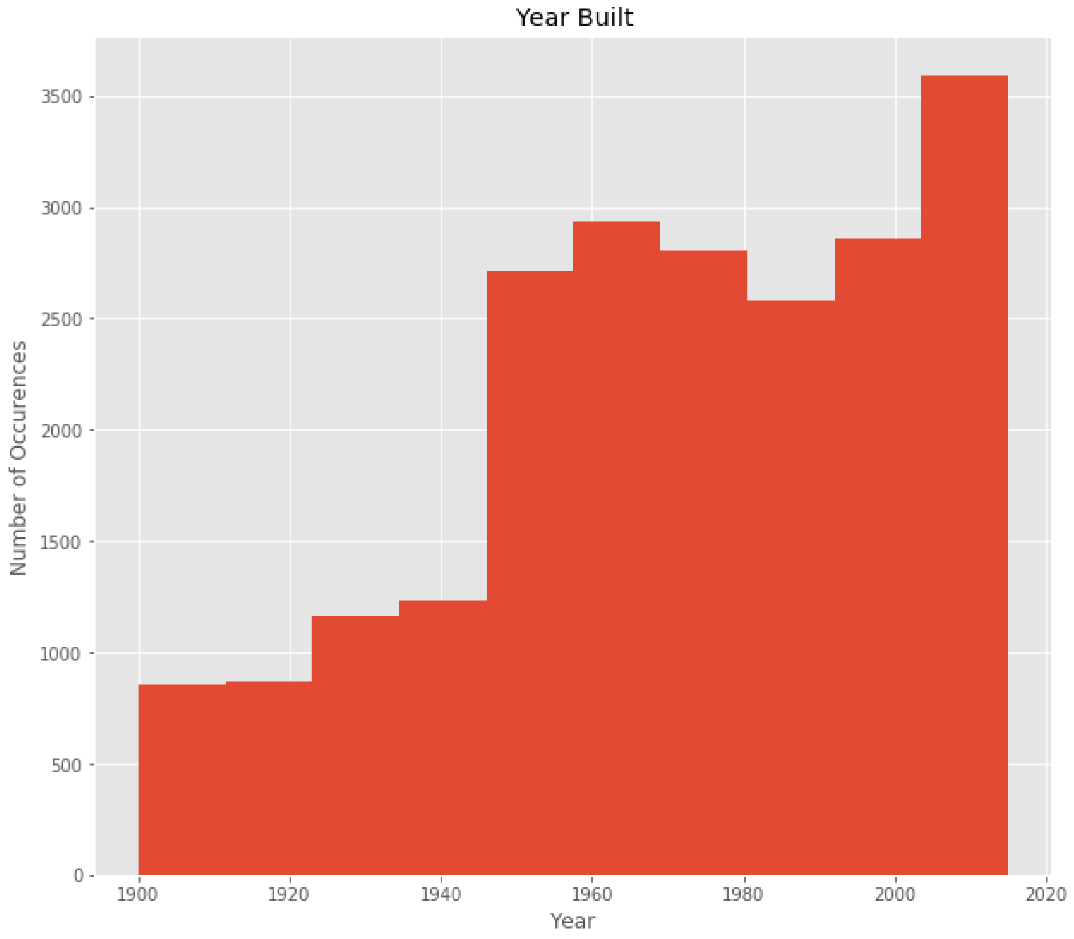
Problem: Dealing with time series data

- Year built has clear affect on price
- Won’t work well in linear model
- Still want to capture affects in linear model



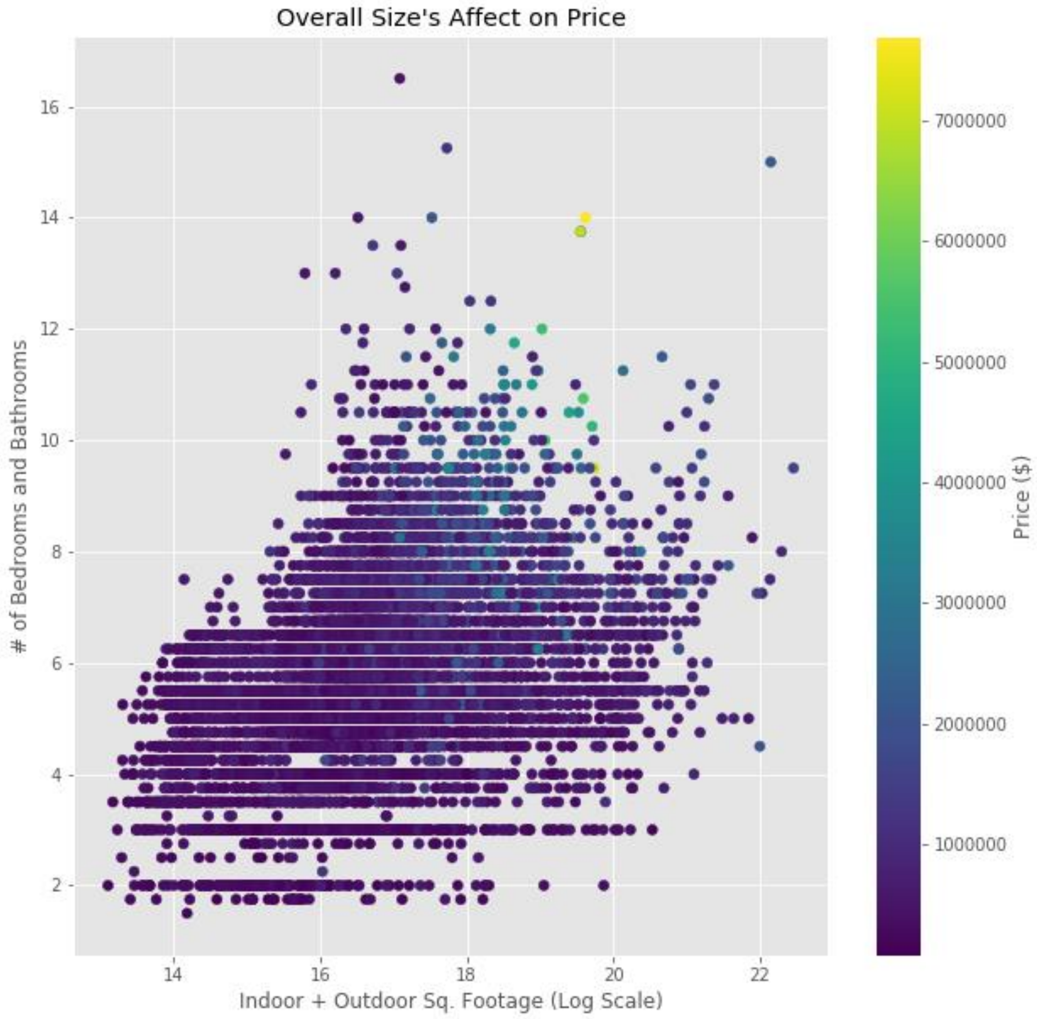
Solution: Binning

- Reduce complexity: Keep only which decade house belongs to
- Create new features representing each bin
- Linear model captures time affect – Performs better



The Model

- Best Features:
  - **Number of Total Rooms & Floors**
  - **Indoor Sq. Footage**
  - **Overall Condition**
  - **Location**
  - **Year Built**
- Achieved R^2 Score of
  - **~0.76 on Test Set**
  - **~0.77 on Training Set**
- Results:
  - Housing prices are volatile, so these results are good for a linear model
  - (See how spread out visual is)



Recommendations - Feature Engineering

- ~**17%** R2 improvement
- All features = statistically significant
- Performs best on new data
- Recommendation: <u>Use model in production</u>

Recommendations - Key Value Drivers

- More expensive homes:
  - Large square footage
  - Waterfront
  - Built past 2009 or from 1900 – 1920
  - Top condition
  - Premium Location (i.e. zipcode 98102)

Recommendations - Things to watch for

- Most expensive homes = most difficult to predict
- Location is everything
  - Identical homes + different neighborhood
    ⬄ **Completely Different Price!**
- # of Rooms **less important** than sq. footage
- Renovation works
  - Home in poor condition
  - Built in 70s – 90s

Recommendations: Improvements

- More Data about location:
  - Nearby Education
  - Parks & Facilities
  - Crime rate
  - Avg. age
  - …