King County Housing Data

Analysis and Observations

- Over 21,000 entries (homes)
- 2014 – 2015
- 21 columns total; analyzed 15 to predict home values.

Methodology

- Evaluate features for distribution of values and quality of data
- Remove those features that don’t add value
- Check for colinear features
- Perform OLS and Multiple Linear Regression analysis
- Perform K-Fold cross validation

Analysis

- Determined that 6 features were not relevant
- Most features not normally distributed: Grade was the best
- Skewed features were primarily Positively skewed due to outliers
- Initial analysis results suggested grade and living space were best predictors
- Mean-normalizing features improved fit
- Train-test results were positive: about 5% difference between them
- K-Fold Cross validation confirmed the results

Key Findings

- Grade was the best predictor, but is subjective as it is human action.
- Of the purely objective features, Living Space is the best predictor.
- Even with Grade, these features only explain (or predict) a little over half of a home's value.
- Eliminating bedroom outliers didn't improve the results.