Minimizing Zestimate® Error

Springboard Career Track - Capstone 1

The problem

Company

Zillow is the leading real estate and rental marketplace dedicated to empowering consumers with data, inspiration and knowledge around the place they call home.

Context

Using Kaggle data from Zillow for three southern California counties, create a model for predicting Zillow's home value estimates (Zestimates®).

Problem statement

Minimize the error (logerror) between estimated home values and the actual sales prices for houses that sold.

Data sources and description

Training data → Properties data → Merged data

Import training data

Over 90K rows containing Parcel ID, log error, and transaction dates for all homes sold in 3 California counties in 2016

Import properties data

Nearly 3MM rows and 58 columns containing data describing the properties for each home in the relevant counties

Merge training and properties data sets and and prepare data for cleaning

- Remove rows with no data
- Remove columns with more than 25% missing values

Data Wrangling

- Dropping columns with more than 25% missing values dropped 28 columns (24 columns remaining)

- Remaining columns were subset and evaluated in logical groups (e.g. Living Area columns)

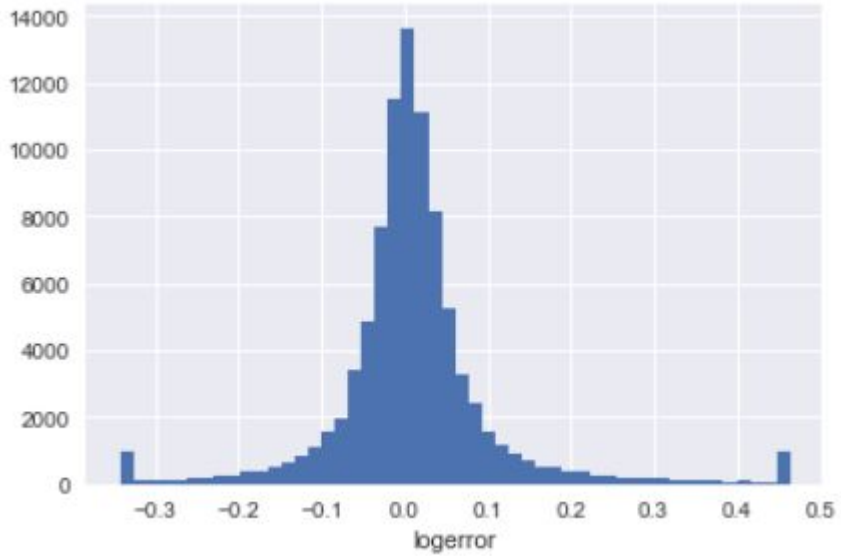
- Tried to impute or otherwise fill in missing values based on known columns or values (e.g. filled in missing zoning county using most common value; sorted by city and filled in missing zip values with forward fill)

Observations

Sales by month numbers are significantly higher starting with month 3 (March) compared with January and February.

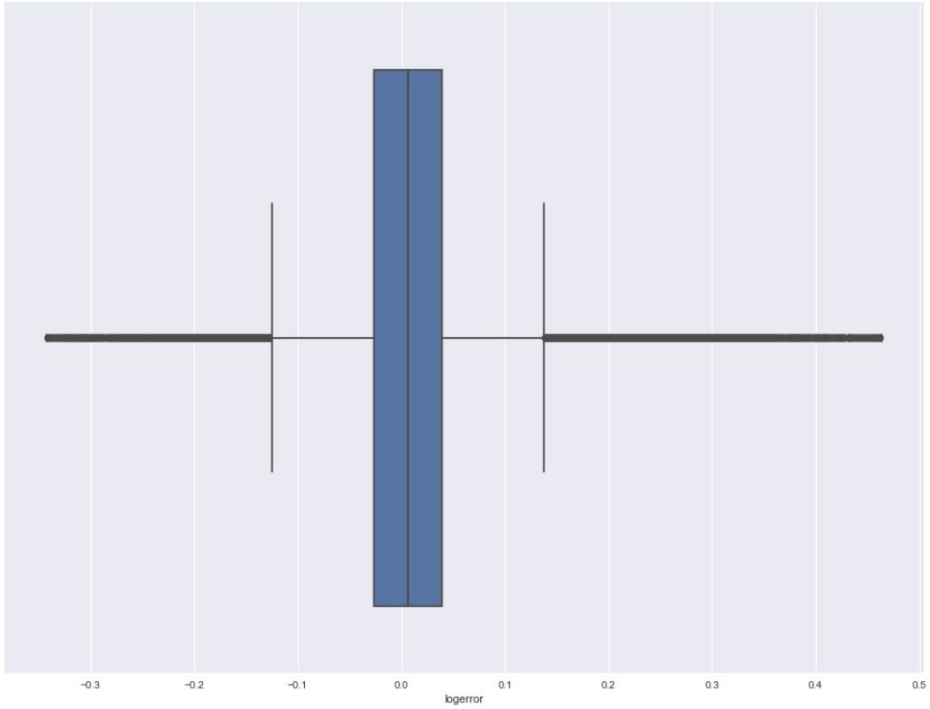
Observations

Logerror values are slightly skewed to the right, with the mean value greater than zero (0.0618) and a longer tail toward the right (outliers collapsed to the 1st and 99th percentile values)



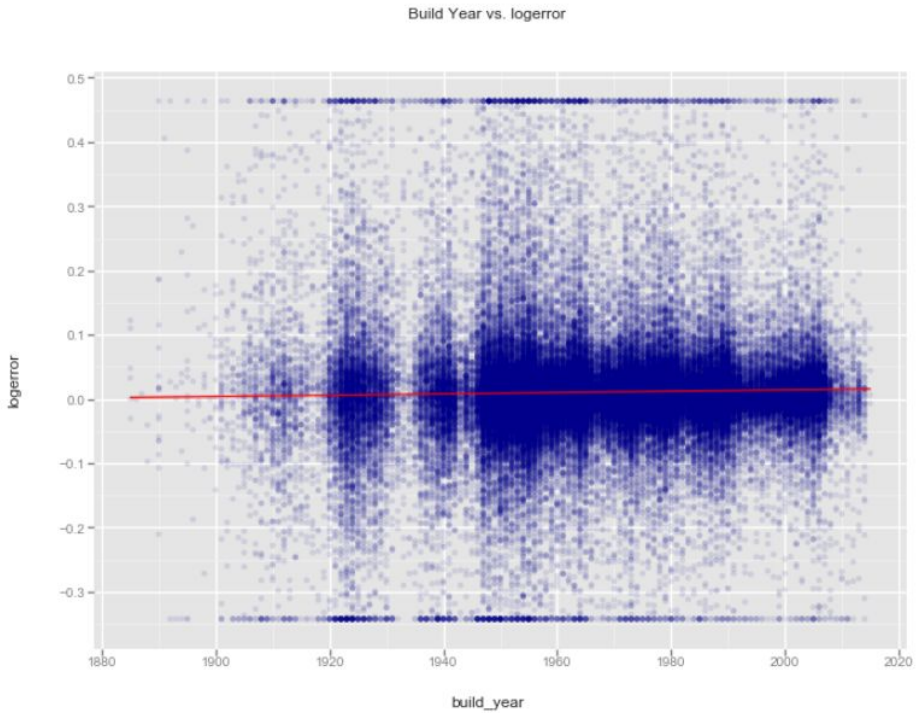
Observations

The median value is 0.0334, with 25% and 75% values of 0.0141 and 0.0714, respectively



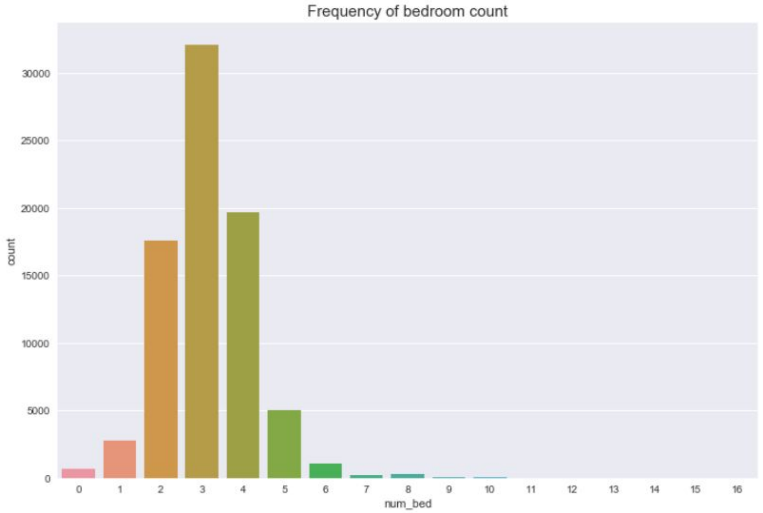
Observations

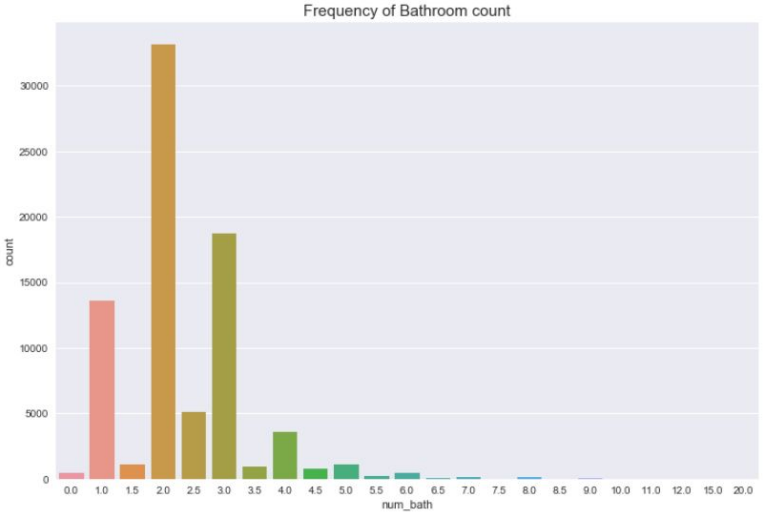
The logerror increases as build_year increases, using a linear model



Observations

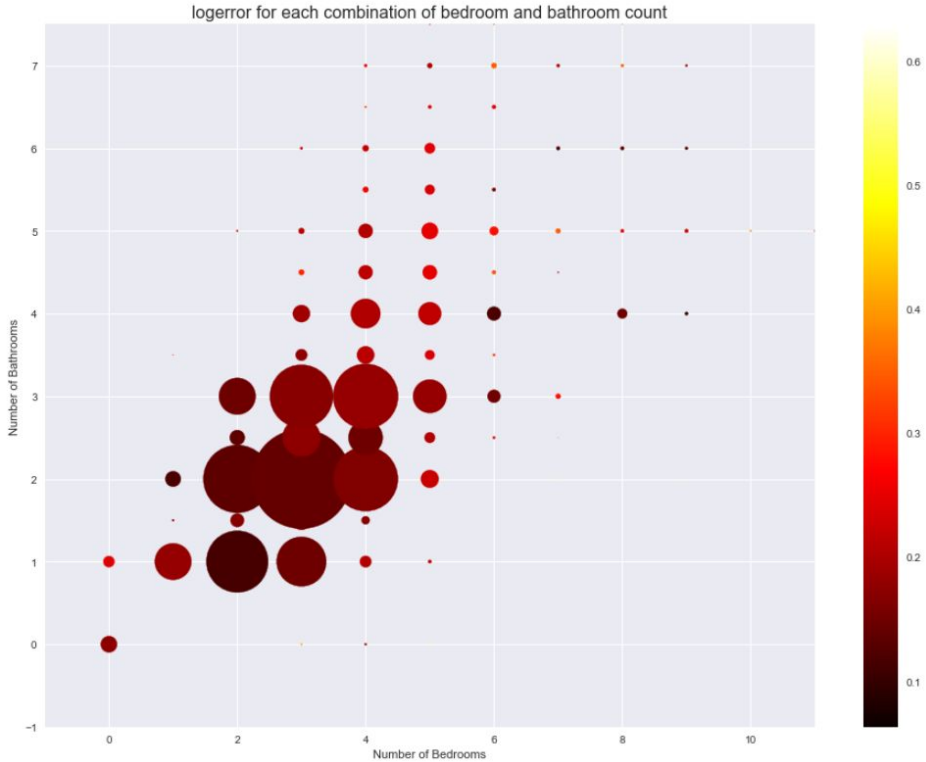
3 BR is the most common number and 2 BA is the most common number. This combination is also the most common





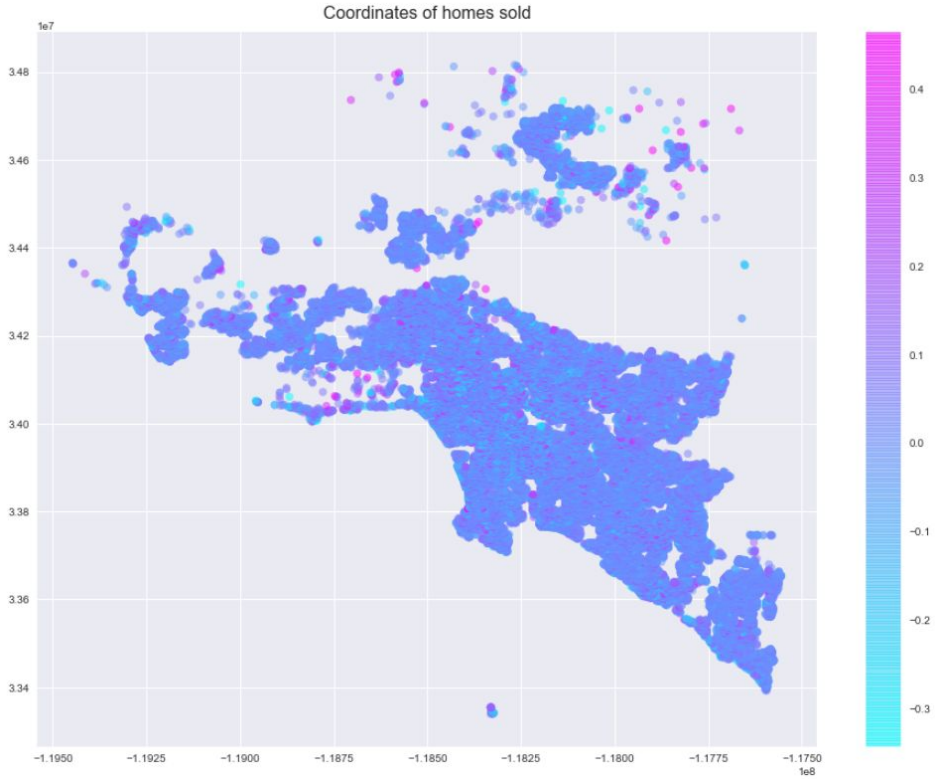
Observations

Generally speaking, the higher the frequency for each combination of bedroom and bathroom count, the lower the absolute value of logerror. The lowest error for houses with fewer than 4 bedrooms and 3 bathrooms appears to be seen for either 2 bedrooms and 1 bath or 1 bedroom and 2 bath (though 2/1 is much more common than 1/2). The logerror for 3/2 homes is on the low side



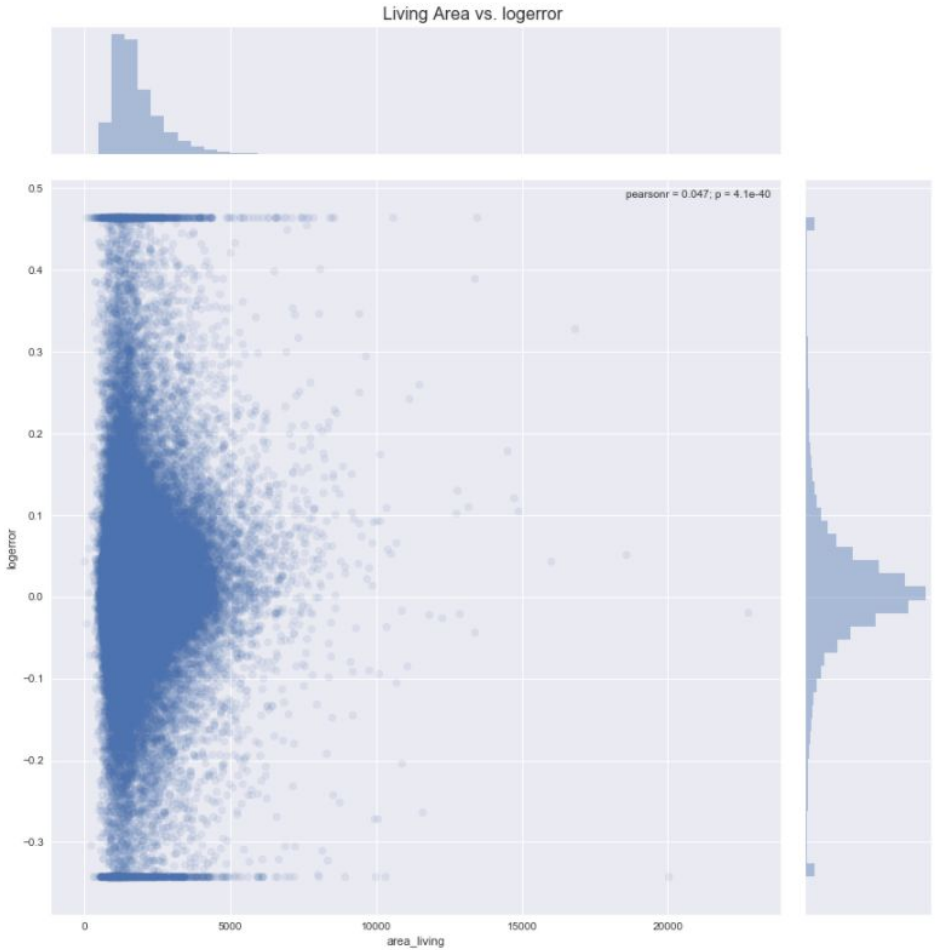
Observations

There do not appear to be any specific areas with higher or lower error rates, they look well distributed.



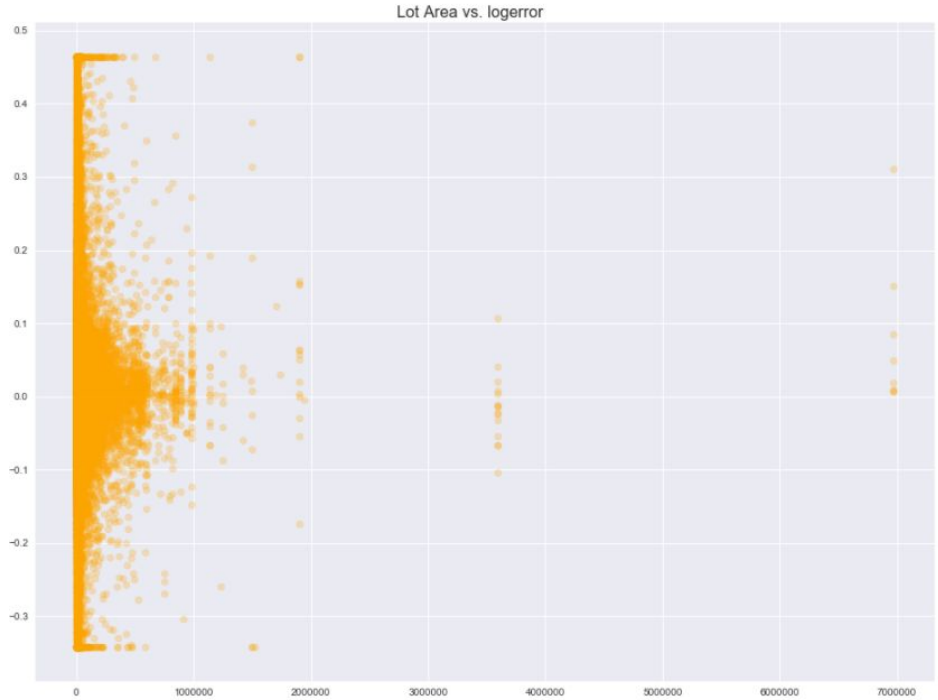
Observations

The range of logerror values becomes smaller as the living area increases. The average living area is 1822 ft^2, with a standard deviation of 952 ft^2.



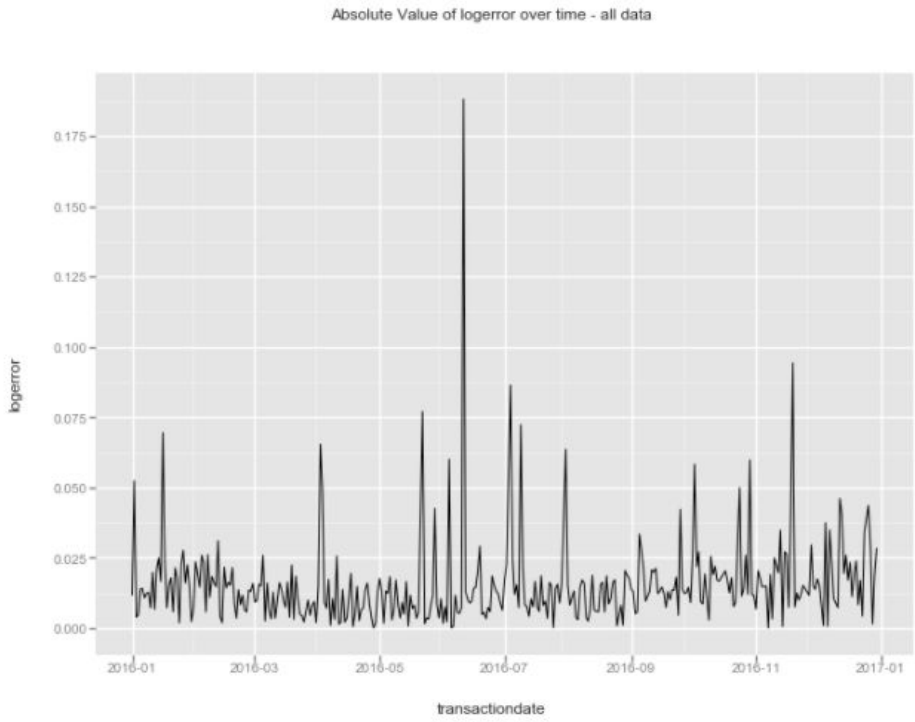
Observations

Likewise, the range of logerror values becomes smaller as the lot area increases



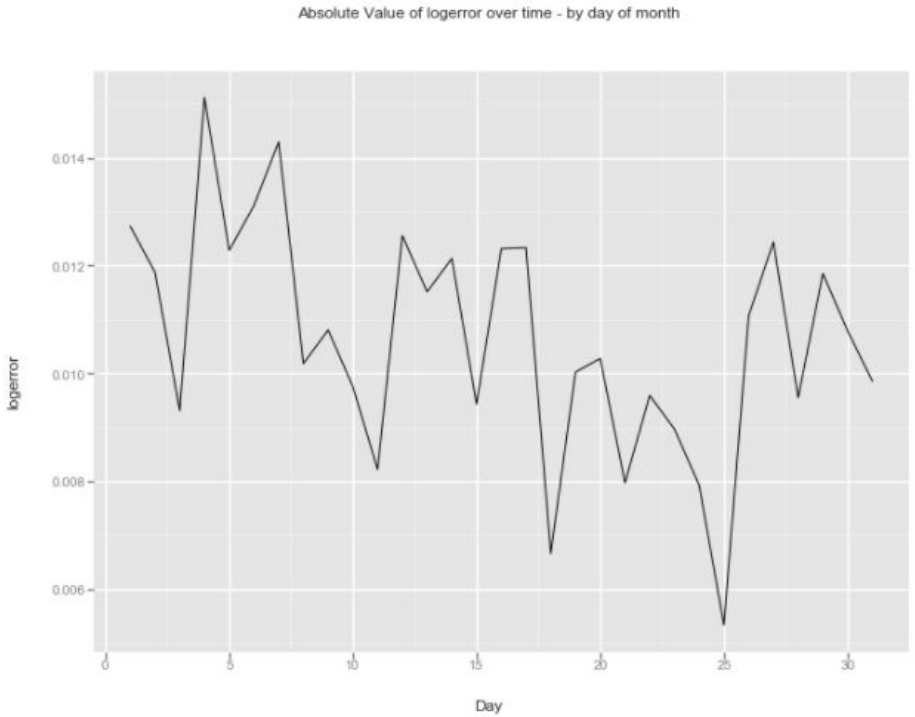
Observations

There are a few spikes in logerror throughout the year, with the largest of those in June 2016 (approx 2x the error of the next largest spike).



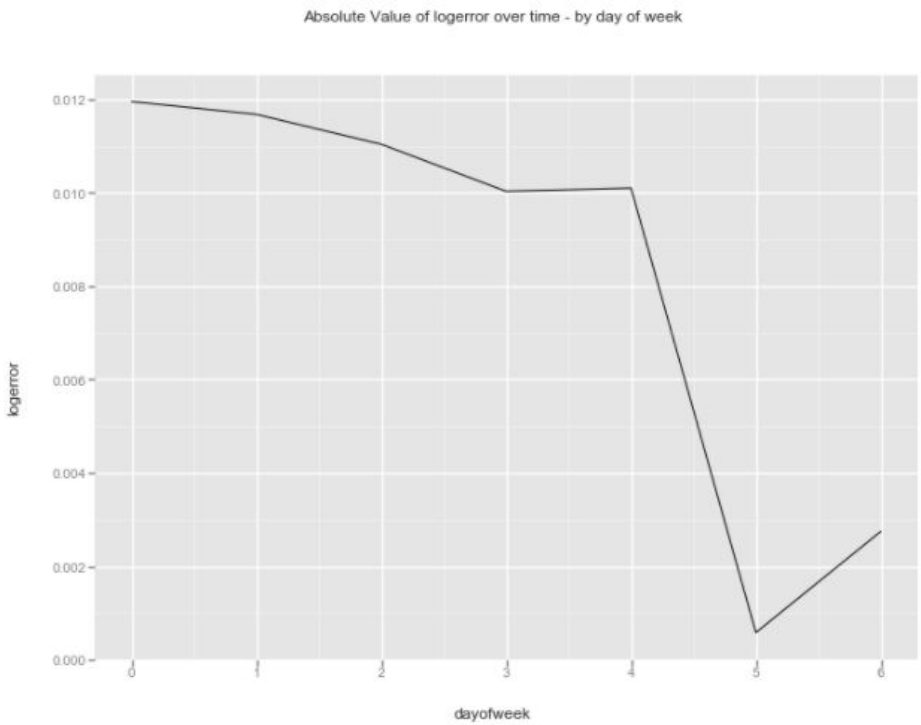
Observations

The average logerror for day of month appears to drop to its lowest point on the 25th of each month.



Observations

The average logerror for day of the week appears to drop to its lowest point on the Saturday of each week, with Sunday being the second lowest. The third lowest, Friday, is 5x the average weekend error.



Inferential Statistics

Correlations:

- Pearson:
  - Highest absolute value correlations to logerror is build_year = 16.1% and county = 8.1%
- Kendall:
  - Top two correlated columns are the same but have lower values (logerror to build_year is approx. 10%)
- Spearman:
  - County now more highly correlated than build_year, but correlation is still just over 10%

Independent t-test for means:

- logerror to number of bedrooms
- logerror to number of bathrooms
- logerror to sum of bedrooms and bathrooms
- logerror to living area
- logerror to lot area

For each of these t-tests, which will be performed as Welch's t-tests due to differing population variances, all null hypotheses are that the means are equal. P-values less than .05 will reject the null hypothesis in favor of the alternate, which is that the means are not equal.

Because all p-values are 0.0, it appears that this test is inadequate. Will proceed with machine learning operations to see if there is any connection between the dependent variable (logerror) and the independent variables.

Predictive Modeling

The following models were created, with some slight variations for Decision Trees and Random Forest.

1. Linear regression
2. Decision Trees
    a. Max depth = 2
    b. Max depth = 5
3. Decision Trees with Adaboost
4. Random Forest
    a. Array with 9 variables
    b. Array with 3 variables

Predictive Modeling

In order from lowest to highest Mean Absolute Error, the errors and methods are:

.0618 - Linear Regression
.0619 - Decision Trees (Max depth = 2)
.0622 - Decision Trees (Max depth = 5)
.0642 - Random Forest (9 variables)
.0672 - Random Forest (3 variables)
.0738 - Decision Trees with Adaboost

This is actually quite counter-intuitive, as looking at the scatterplots, it would have appeared that the Random Forest regressions would have had the lowest error and the Linear Model would have had the highest.