# Overview

The goal of the project is to build a model to predict “logerror” which is the difference between the Zillow’s estimated home value, Zestimate, and the actual sale price.

*logerror=log(Zestimate)−log(SalePrice)*

It is to improve the Zestimate’s residual error by predicting where zestimates will do good or bad. When we want to improve existing model, modeling errors can be a good way to find areas to improve the existing model.

## Data

The data used in the project has been provided from Zillow through Kaggle.com. The data is found at :

https://www.kaggle.com/c/zillow-prize-1

The following two files were used in the project.

1. **properties_2016.csv**: The full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016. The dataset covers a wide range of information, including 59 features such as the building framing type, area of the lot in square feet, zip code , total property tax and etc.
2. **train_2016.csv**: all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016. It contains parcel ID , transaction date and calculated log error .

The two dataset were joined by “parcel ID” and produced the dataset with 90275 transactions and 59 features. The data consistes of all the transations before

# Data Wrangling

## Data Cleaning

- Duplication : I explored training data. 125 duplicated data for 2016 and 199 duplicated data for 2017 data were found. However, it meant they were transitioned for more than twice for a year. So, I didn’t delete any duplication.

- Negative values: Also, I checked if there were any negative numbers for each column. Two columns, logerror and longitude, have negative values which are reasonable to have for them.

- Unusual Object : We have 5 columns which are objective type. Each column does not have unusual values, for example “?” , “$”

## Missing Values

Let's check how many missing values each column has. I found that 47 columns out of 60 columns have missing values and 18 columns among them have more than 95% of missing values.

Let's explore how missing values were treated.

| | Column name | Describtion | Missing | Missing Values |
|---|---|---|---|---|
| 1 | buildingclasstypeid | The building framing type | 99.98% | Deleted the column because only 16 cells out of 90275 cells are not missing and all with the same value 4. Rest of data, 90259 are missing for random. |
| 2 | finishedsquarefeet13 | Perimeter living area | 99.96% | Delete the column because every build must have living area and most of them are missing. |
| 3 | basementsqft | Finished living area below ground level | 99.95% | Filled with 0 because every building does not have living area below ground leve, missing could mean building does not have partial living room. |
| 4 | storytypeid | Type of floors in a multi-story house | 99.95% | Deleted the column because every building should have a type of floor and most of them are missing . |
| 5 | yardbuildingsqft26 | Storage shed/building in yard | 99.89% | Filled with 0 because missing value can mean it doesn't have storage in yard |
| 6 | fireplaceflag | Is a fireplace present in this home | 99.75% | Filled with False because missing value means it does not have a fireplace. |
| 7 | architecturalstyletypeid | Architectural style of the home | 99.71% | Deleted the column because every buildling has its architectural style and most of them are missing. |
| 8 | typeconstructiontypeid | type of construction used to construct the home | 99.67% | Deleted the column because every building has its type of construction material and most of them are missing |
| 9 | finishedsquarefeet6 | Base unfinished and finished area | 99.53% | Deleted the column because everyg home should have base area and most of them are missing |
| 10 | decktypeid | Type of deck present on parcel | 99.27% | Deleted the column because non-missing cells have the same value, 66 and most of them are missing |
| 11 | poolsizesum | Total square footage of all pools on property | 98.93% | Deleted the column because it is missing randomly and most of them are missing |
| 12 | pooltypeid10 | Spa or Hot Tub | 98.71% | Deleted the column because it is missing randomly and most of them are missing |
| 13 | pooltypeid2 | Pool with Spa/Hot Tub | 98.67% | Deleted the column because it is missing randomly and most of them are missing |
| 14 | taxdelinquencyflag | Property taxes for this parcel are past due as of 2015 | 98.02% | Filled with Y because all non-missing values are "N" |
| 15 | taxdelinquencyyear | Year for which the unpaid propert taxes were due | 98.02% | Filled with 0 because the missing cells are the same as the previoius column. |
| 16 | hashottuborspa | Does the home have a hot tub or spa | 97.38% | Filled with False because all non-missing cells are "True" |
| 17 | yardbuildingsqft17 | Patio in yard | 97.07% | Filled with 0 because not every building has a patio in yard. |
| 18 | finishedsquarefeet15 | Total area | 96.05% | Deleted the column because every place should have total area and most of them are missing |
| 19 | finishedfloor1squarefeet | Size of the finished living area on the first floor of the home | 92.41% | Deleted the column because most place has living area and most of cells are missing |
| 20 | finishedsquarefeet50 | Size of the finished living area on the first floor of the home | 92.41% | Deleted the column because it is the repeat of the previous column. |
| 21 | fireplacecnt | Number of fireplaces in a home (if any) | 89.36% | Filled with 0 because not every building has fireplace. |

| | | | | |
|---|---|---|---|---|
| 22 | threequarterbathnbr | Number of 3/4 bathrooms in house (shower + sink + toilet) | 86.70% | Filed with 0 because no all home has 3/4 bathrooms |
| 23 | pooltypeid7 | Pool without hot tub | 81.50% | Deleted the column because not every home has a pool and most of them are missing. |
| 24 | poolcnt | Number of pools on the lot (if any) | 80.17% | Filled with 0 because no every home has pool. |
| 25 | numberofstories | Number of stories or levels the home has | 77.21% | Deleted the column because every home should have a number of levels and most of them are missing. |
| 26 | airconditioningtypeid | Type of cooling system present in the home (if any) | 68.12% | Filled with 0 because not all home have a cooling system. |
| 27 | garagetotalsqft | Total number of square feet of all garages on lot including an attached garage | 66.84% | Deleted the column. Missing might mean no garage, but there are non-missing cells with 0. |
| 28 | garagecarcnt | Total number of garages on the lot including an attached garage | 66.84% | Deleted the column with the same reason with the previous. |
| 29 | regionidneighborhood | Neighborhood in which the property is located | 60.11% | Deleted the column because it is missing randoly. |
| 30 | heatingorsystemtypeid | Type of home heating system | 37.88% | Filed with 0 because not all home have heating system. |
| 31 | buildingqualitytypeid | Overall assessment of condition of the building | 36.46% | Filled with mean because all home have overall assessment. |
| 32 | propertyzoningdesc | Description of the allowed land uses (zoning) for that property | 35.41% | Filled with "Missing" to treat missing values as another class |
| 33 | unitcnt | Number of units the structure is built | 35.36% | Filled with 1 because 1 is the most frequent value. |
| 34 | lotsizesquarefeet | Area of the lot in square feet | 11.24% | Filled with mean because all home has are of the lot. |
| 35 | finishedsquarefeet12 | Finished living area | 5.18% | Filled with mean because all home has living area. |
| 36 | regionidcity | City in which the property is located (if any) | 2.00% | Filled with the most frequent value because all home is located in city. |
| 37 | fullbathcnt | Number of full bathrooms present in home | 1.31% | Filed with 0 because missing might mean home does not have full bathroom. |
| 38 | calculatedbathnbr | Number of bathrooms in home including fractional bathroom | 1.31% | Filed with 0 because missing might mean home does not have full bathroom. |
| 39 | yearbuilt | The Year the principal residence was built | 0.84% | Filled with mean because all home have the year built in. |
| 40 | calculatedfinishedsquarefeet | Calculated total finished living area of the home | 0.73% | Filled with mean because most home have living room |
| 41 | censustractandblock | Census tract and block ID combined | 0.67% | Filled with most frequent value because every home has it's value |
| 42 | structuretaxvaluedollarcnt | The assessed value of the built structure on the parcel | 0.42% | Filled with mean because every home has the assessed value |
| 43 | regionidzip | Zip code in which the property is located | 0.04% | Filled with the most frequent value because every home has zip code. |
| 44 | taxamount | The total property tax assessed for that assessment year | 0.01% | Filed with mean because every home has property tax. |
| 45 | taxvaluedollarcnt | The total tax assessed value of the parcel | 0.00% | Filed with mean because every home has property tax. |
| 46 | landtaxvaluedollarcnt | The assessed value of the land area of the parcel | 0.00% | Filled with mean because every home have assesed value. |
| 47 | propertycountylandusecode | County land use code i.e. it's zoning at the county level | 0.00% | Filled with the most frequent value. |

## Categorical values to dummy variables

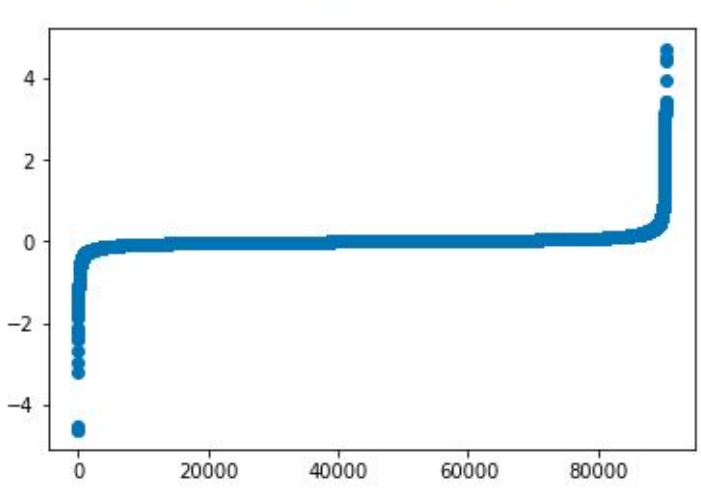
To use variables in the model, we need to convert categorical data to dummies variables. Also, some categorical data have too many columns, so we need to group

columns for each categorical values.

| | Column name | Describtion | No of Columns | How to group |
|---|---|---|---|---|
| 1 | propertyzoningdesc | Description of the allowed land uses (zoning) for that property | 1996 | The most frequent value is "LAR1" and has frequency of 7678, 8.5%. We would use values with more than 1%. Other values with less than 1% is too small. We would label values with less than 1% to "Others" |
| | regionidcity | City in which the property is located | 177 | label values with less than 1% to "Others" |
| 3 | regionidzip | Zip code in which the property is located | 388 | The most frequent value is 97319 and it is 1% of entire data. All variables consist less than 1% of entire data. Therefore, group categorical data to 4 groups using percentile. |
| 4 | propertycountylandusecode | County land use code i.e. it's zoning at the county level | 77 | The most frequent is "0100" with 34%. There are 12 values that each value consists of more than 1% of entire datay. I would change any values with less than 1% of the entire data to others. |
| | | | | |

## Outliers

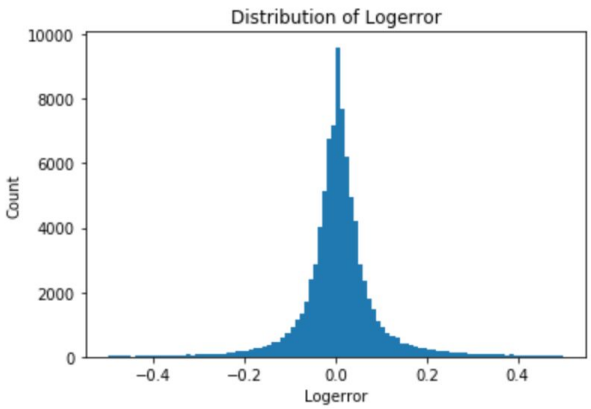
Let's draw a scatter plot on "logerror", then we can find that there are some outliers at the end of both sides. Our task in the project is to find where the zillow algorithm fails. These outliers means where the zillow algorithm fails the most. Thus, I will leave outliers just like that.

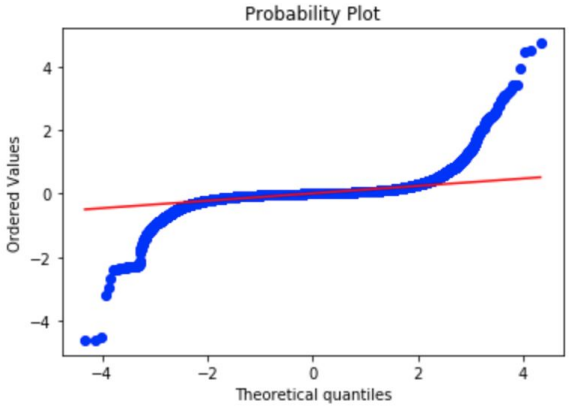



# Data Storytelling

## Distribution of Logerror

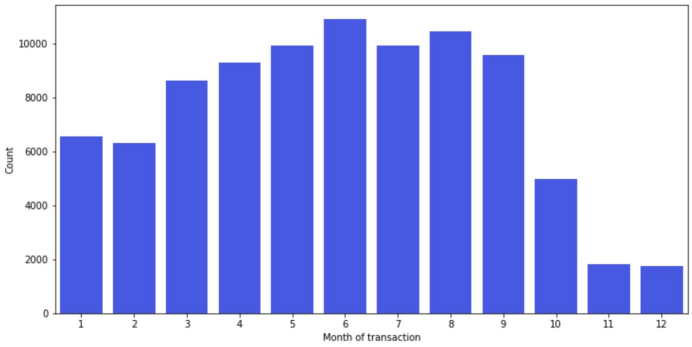
It seems like the distribution of logerror follows a normal distribution by checking QQ plot.



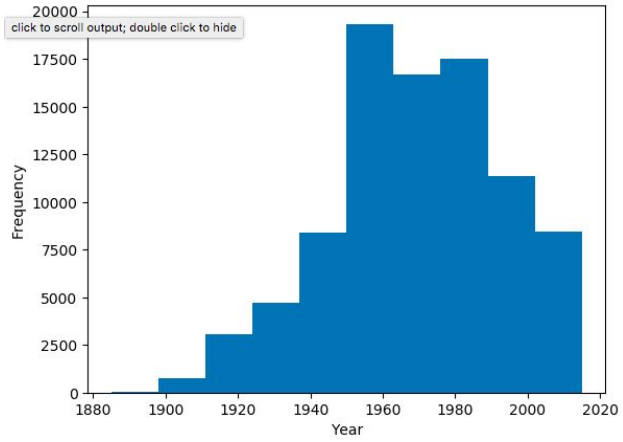

## Transaction Dates

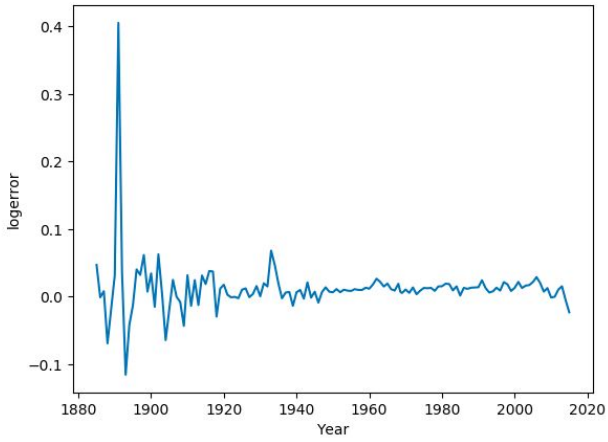
Let's check the distribution of transaction dates, there are fewer transactions after October. As the data consists of all the transactions before October 15, 2016, plus some of the transactions after October 15, 2016, there are fewer transactions after October.




## Built year

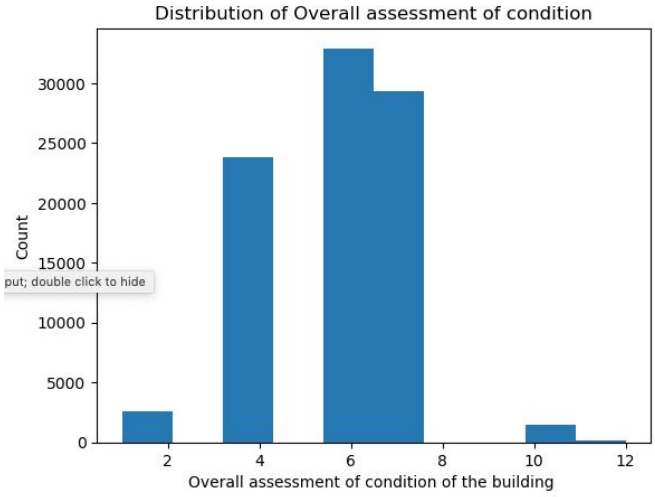
After observing plot for density of built year, we can find that most houses, 59.37%, are built between 1950 and 1990. Logerror is getting smaller with newer houses. Zestimate predicts home value better with newer homes.



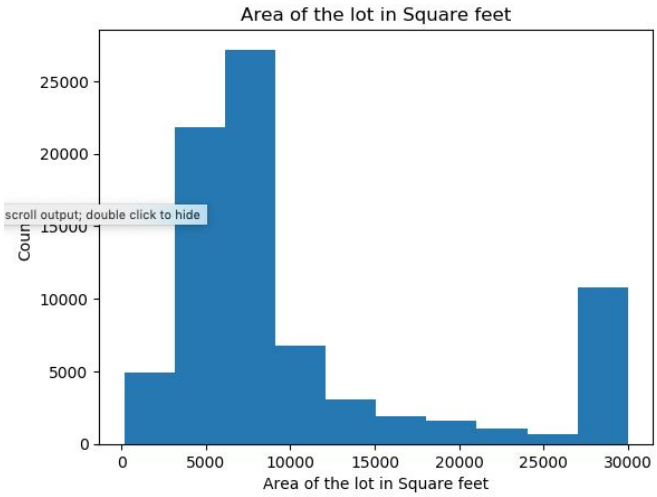

## Overall Assessment of condition

"buildingqualitytypei" is overall assessment of condition of the building rates from best (lowest) to worst (highest) between 1 to 12. The mean is 5.56 and 68.9% of data are between 5 to 8.



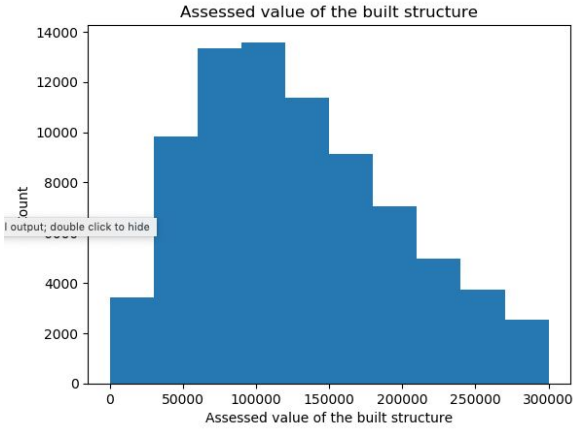

## Size of lot

"lotsizesquarefeet" is area of the lot in square feet. The mean of lot size is 29110. 25% of the data is between 5962 and 7570. The maximum is 6971010.



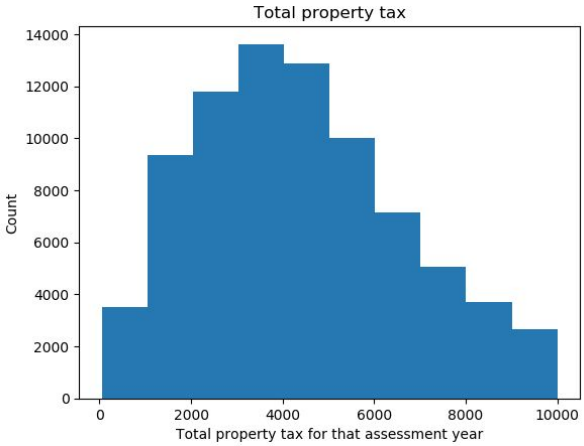

## The assessed value of the built structure

"structuretaxvaluedollarcnt" is the assessed value of the built structure. The mean value of it is 180093 and maximum is 9948100




## Tax

"taxamount" is the total property tax assessed for that assessment year. The mean is 5983.




# Inferential Statistics

Let's check correlations of each variables to "logerror" to see how variables are related.

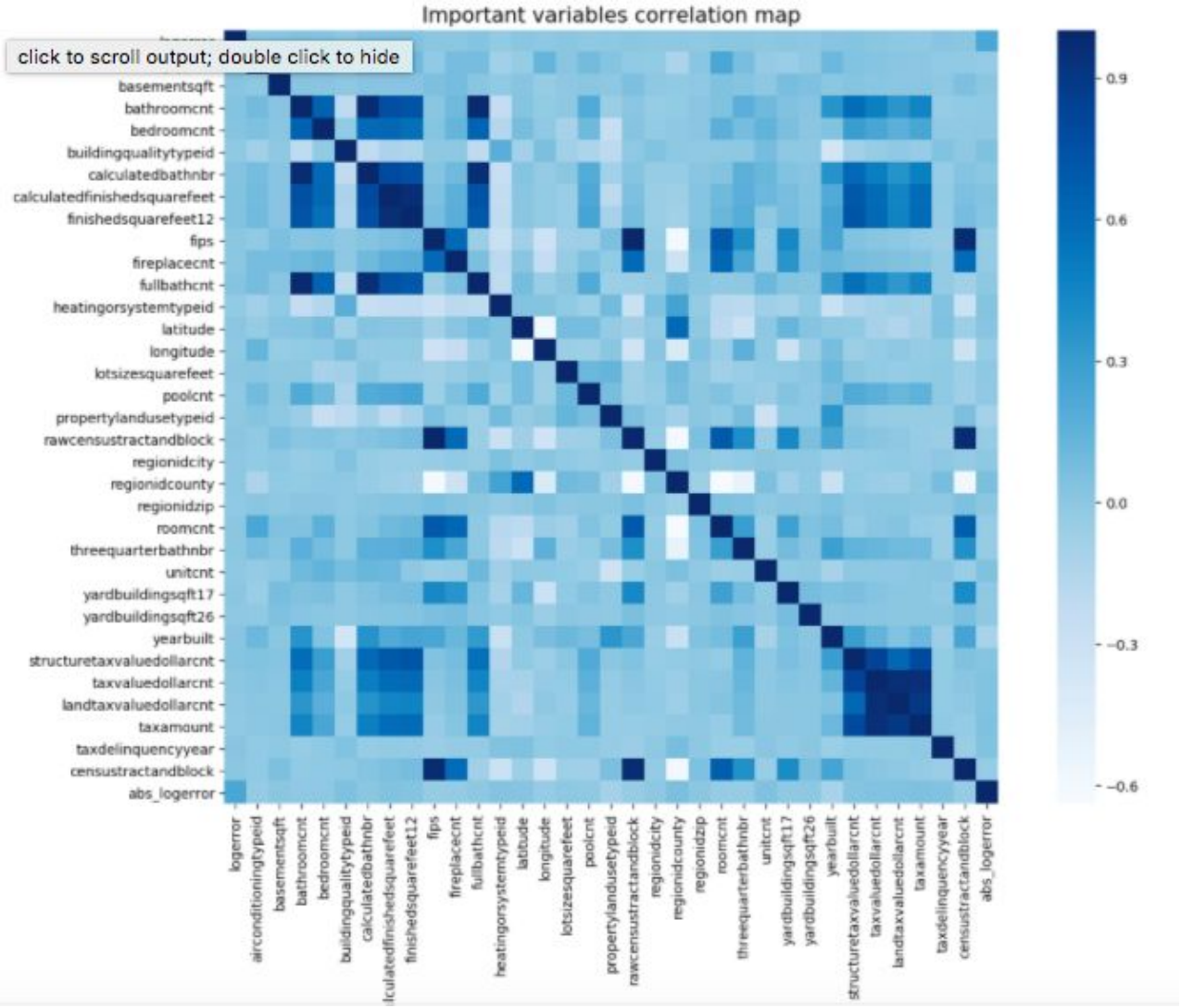
Correlation between target variable, logerror, and dependent variables are all weak. They are between 0.237380 and -0.018009.

| Name | Coeff |
|---|---|
| abs_logerror | 0.23738 |
| finishedsquarefeet12 | 0.039248 |
| calculatedfinishedsquarefeet | 0.038341 |
| calculatedbathnbr | 0.028788 |
| bathroomcnt | 0.027889 |
| fullbathcnt | 0.027571 |
| bedroomcnt | 0.025467 |
| structuretaxvaluedollarcnt | 0.021935 |
| taxdelinquencyyear | 0.018107 |
| yearbuilt | 0.017089 |
| basementsqft | 0.009019 |
| rawcensustractandblock | 0.008376 |
| fips | 0.008363 |
| fireplacecnt | 0.007746 |
| taxvaluedollarcnt | 0.006508 |
| roomcnt | 0.00576 |
| threequarterbathnbr | 0.00549 |
| airconditioningtypeid | 0.005404 |
| latitude | 0.004915 |
| lotsizesquarefeet | 0.004612 |
| censustractandblock | 0.004495 |
| yardbuildingsqft17 | 0.002497 |

| | |
|---|---|
| propertylandusetypeid | 0.001003 |
| regionidcounty | 0.000341 |
| yardbuildingsqft26 | -0.000846 |
| regionidcity | -0.002342 |
| landtaxvaluedollarcnt | -0.003051 |
| longitude | -0.003432 |
| unitcnt | -0.003447 |
| regionidzip | -0.006487 |
| taxamount | -0.006671 |
| buildingqualitytypeid | -0.00788 |
| poolcnt | -0.008983 |
| heatingorsystemtypeid | -0.018009 |

Let's check correlations between pairs of independent variables. We can find that there are 2 clusters on the heat map below. The first cluster on the top left shows that variables about sizes of houses such as bathroom size or bedroom size and total square feets are stronly related. The second cluster on the bottom right tells us that variables about taxes are related to each other. Alse, we can observe from the right top cluster that variables about sizes of houses are weakly related to variables about taxes. It is reasonable because the bigger a house is, the more expensive the property is resulting the more taxes. However, the price of house is not only resulted from the size of house. So correlation is not strong among them.

There are not variables which can be particularly significant in terms of predicting logerror based on correlation. Also, there are strong multicollinearity between dependent variables. Therefore, a linear regression is not suitable for the model because of multicollinearity.




# Machine Learning

## Linear Regression

Training data was randomly chosen from 75% of entire data 50 times to get mean squared error. The mean squared error was 0.02603 which were the average of 50 mean squared errors. R squared is 0.0052

The top 10 properties which have large impacts on logerrors were like the table below. The 3 properties which have the most impact are "airconditioningtypeid_3.0", "propertylandusetypeid_263.0" and "propertylandusetypeid_31.0". "airconditioningtypeid_3.0" is the type of cooling system present in the home which is Evaporative Cooler. "propertylandusetypeid_263.0" and "propertylandusetypeid_31.0" are the type of land use the property is zoned for. From the table below, "propertylandusetypeid" and "propertycountylandusecode" have large impact on "Logerror".

| name | coeff | |
|---|---|---|
| **102** | propertylandusetypeid_263.0 | 0.123889 |
| **103** | propertylandusetypeid_264.0 | 0.046871 |
| **88** | propertycountylandusecode_1 | 0.040308 |
| **108** | propertylandusetypeid_275.0 | 0.039954 |
| **92** | propertycountylandusecode_122 | 0.037242 |
| **93** | propertycountylandusecode_34 | 0.037155 |
| **71** | regionidcity_47568.0 | 0.025231 |
| **4** | calculatedbathnbr | 0.018124 |
| **43** | heatingorsystemtypeid_18.0 | 0.017372 |
| **34** | heatingorsystemtypeid_1.0 | 0.016581 |

## Random Forest

To find the best fitted random forest model, grid search is used. possible combination of options were applied to find the better model. From grid search, the model with max_depth of 5, min_samples_spli of 20 and n_estimators of 30 was selected.

The MSE for the model was 0.02580. So random forest is a good to predict logerror.

Also the top five variables with large coeffients are in the below. As we can see below, variables related to tax or living area had high impact on logerror.

| name | description | coeff |
|---|---|---|
| taxamount | tax amount | 0.158339 |

| | | |
|---|---|---|
| longitude | longitude | 0.128765 |
| finishedsquarefeet12 | finished living area | 0.119519 |
| calculatedfinishedsquarefeet | total finished living room area | 0.112954 |
| taxvaluedollarcnt | total tax assessed value of the parcel | 0.092970 |

## Gradient Boosting Regressor

To find the best parameters for Gradient Boosting Regressor, Grid search was used. After the best parameter chosen. The mse is  0.02488.

| name | coeff | coeff |
|---|---|---|
| taxamount | taxamount | 0.126420 |
| structuretaxvaluedollarcnt | The assessed value of the built structure on the parcel | 0.113069 |
| taxvaluedollarcnt | total tax assessed value of the parcel | 0.104929 |
| calculatedfinishedsquarefeet | total finished living room area | 0.085786 |
| finishedsquarefeet12 | finished living area | 0.07742 |

## Lasso

Let's try Lasso. To find the alpha for Lasso try many possible variables, 0,0.0001,0.001, 0.01,0.1,0.5,1,2,3,4 for alpha, then choose the best variables. The best alpa which gives lowest mse is 0.001 with mse of 0.02896

11 variables were chosen for Lasso, but coefficients for each chosen variables are low. Therefore Lasso is not a good model to predict logerror

| name | description | coeff |
|---|---|---|
| taxvaluedollarcnt | The total tax assessed value of the parcel | 1.010697e-07 |
| lotsizesquarefeet | Area of the lot in square feet | 3.656718e-09 |
| latitude | latitude | 2.463666e-09 |
| longitude | longitude | 1.907122e-11 |
| structuretaxvaluedollarcnt | The assessed value of the built structure on the parcel | -4.335424e-08 |
| landtaxvaluedollarcnt | The assessed value of the land area of the parcel | -8.562764e-08 |
| taxamount | The total property tax assessed for that assessment year | -2.273446e-06 |

# Conclusion

| Model | MSE |
|---|---|
| Linear Regression | 0.02603 |
| Random Forest | 0.02580 |
| Gradient Boosting Regressor | 0.02488 |
| Lasso | 0.02896 |

As we see on the above, the Gradient Boosting Regressor has the lowest MSE. MSE is the measure of how predicted values are different from the actual values, the smaller MSE the better. So Gradient Boosting Regressor is the model I chose. The variables which have larger impact on logerror are related to tax, living area and total assessment value. So we need to improve Zillow's estimated home value, Zestimate in those areas.

The next step for this project is to gather more data as current data is limited to 2016 with fewer data on October.