~~$350,000~~

$300,000

PREDICTING LOGERROR IN ZILLOW REAL ESTATE PRICING PREDICTIONS

-PETER GIERKE

Background on Zillow

- *Zillow has data on 110 million homes across the United States, not just those homes currently for sale.*
- *Zillow determines an estimate ("Zestimate," pronounced "ZEST-imate") for a home based on a range of publicly available information, including sales of comparable houses in a neighborhood.*
- *In 2007, The Wall Street Journal studied the accuracy of Zillow's estimates and found that they "often are very good, frequently within a few percentage points of the actual price paid. But when Zillow is bad, it can be terrible.*
- *Using data published on the Zillow website, the typical Zestimate error in the United States in July 2016 was $14,000.*

https://en.wikipedia.org/wiki/Zillow

Background on competition

- *Zillow is looking to improve the accuracy of Zestimates*
- *Target for prediction is the log error as described:*
  - *logerror=log(Zestimate)−log(SalePrice)*
- *Goal is to predict logerror of 6 individual timepoints:*
  - *October, November and December of 2016 and October, November and December of 2017*
- *Submission Accuracy is evaluated on MAE (Mean Absolute Error)*

Dataset -Data

- Data
  - 2,985,217 samples
  - 58 columns of data.
    - 1 of type int
    - 5 of type object
    - 52 of type float float
  - Large portion of the columns missing more than 80% of the data.

parcelid	0.000000
logerror	0.000000
transactiondate	0.000000
airconditioningtypeid	68.306703
architecturalstyletypeid	99.712590
basementsqft	99.952649
bathroomcnt	0.590237
bedroomcnt	0.590237
buildingclasstypeid	99.982381
buildingqualitytypeid	36.831441
calculatedbathnbr	1.891841
decktypeid	99.275418
finishedfloor1squarefeet	92.450254
calculatedfinishedsquarefeet	1.318122
finishedsquarefeet12	5.742696
finishedsquarefeet13	99.963661
finishedsquarefeet15	96.075365
finishedsquarefeet50	92.450254
finishedsquarefeet6	99.536400
fips	0.590237
fireplacecnt	89.420885
fullbathcnt	1.891841
garagecarcnt	67.033729
garagetotalsqft	67.033729
hashottuborspa	97.395690
heatingorsystemtypeid	38.245367
latitude	0.590237
longitude	0.590237
lotsizesquarefeet	11.767297
poolcnt	80.287630
poolsizesum	98.932949
pooltypeid10	98.721521
pooltypeid2	98.674169
pooltypeid7	81.613461

propertycountylandusecode	0.591338
propertylandusetypeid	0.590237
propertyzoningdesc	35.786414
rawcensustractandblock	0.590237
regionidcity	2.575679
regionidcounty	0.590237
regionidneighborhood	60.344011
regionidzip	0.628778
roomcnt	0.590237
storytypeid	99.952649
threequarterbathnbr	86.775831
typeconstructiontypeid	99.670745
unitcnt	35.742366
yardbuildingsqft17	97.086256
yardbuildingsqft26	99.895387
yearbuilt	1.422735
numberofstories	77.348559
fireplaceflag	99.755536
structuretaxvaluedollarcnt	1.008688
taxvaluedollarcnt	0.591338
assessmentyear	0.590237
landtaxvaluedollarcnt	0.591338
taxamount	0.596844
taxdelinquencyflag	98.036581
taxdelinquencyyear	98.036581
censustractandblock	1.256456

Data -Target

- *3 columns*
  - *1 nominal (parcel ID)*
  - *1 continuous (logerror) *competition target*
  - *1 interval (transaction date)*
- *90811 samples*

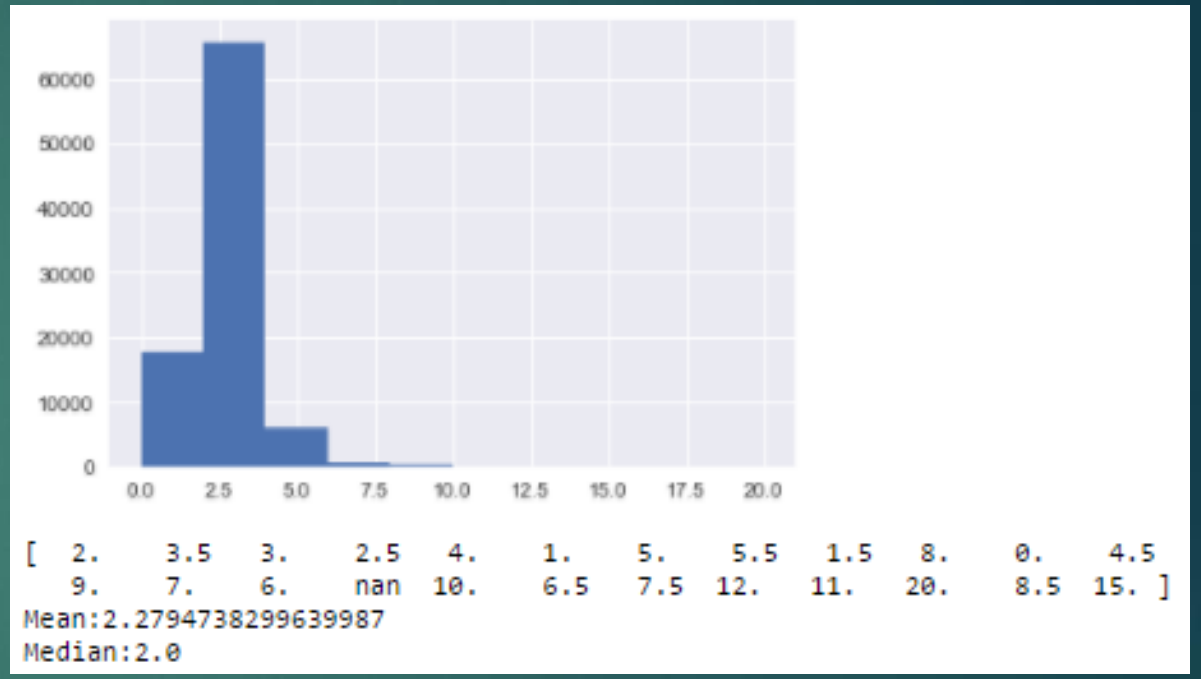
	parcelid	logerror	transactiondate
0	11016594	0.0276	2016-01-01
1	14366692	-0.1684	2016-01-01
2	12098116	-0.0040	2016-01-01
3	12643413	0.0218	2016-01-02
4	14432541	-0.0050	2016-01-02

Data Cleaning and Feature Engineering

- *Merge data onto target dataset to start working with features for available target data.*
- *Drop data where more than 20% of the column is NaN. Leaves us with 26 columns.*

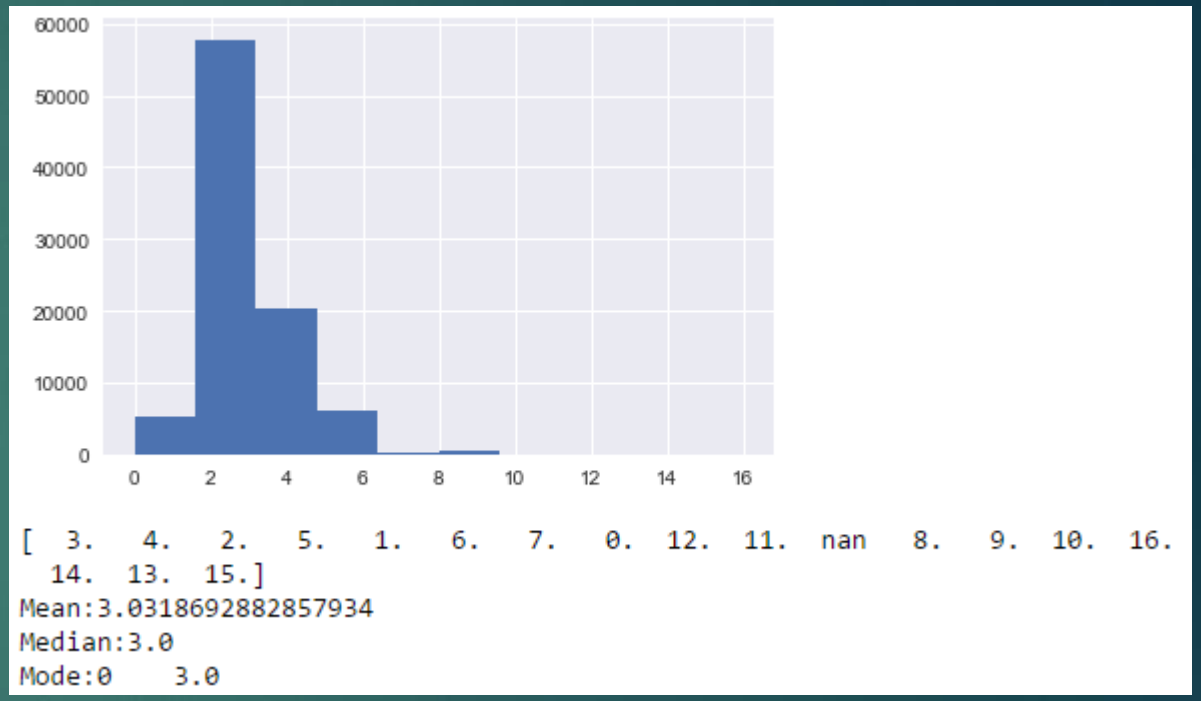
Data Cleaning and Feature Engineering

- *bathroomcnt:*
- *(Number of bathrooms in home including fractional bathrooms)*
- *0.59% NaN*
- *Mean is 2.27, which is not a possible datapoint for bathroom count.*
- *Fill NaN values with median value of 2.0*



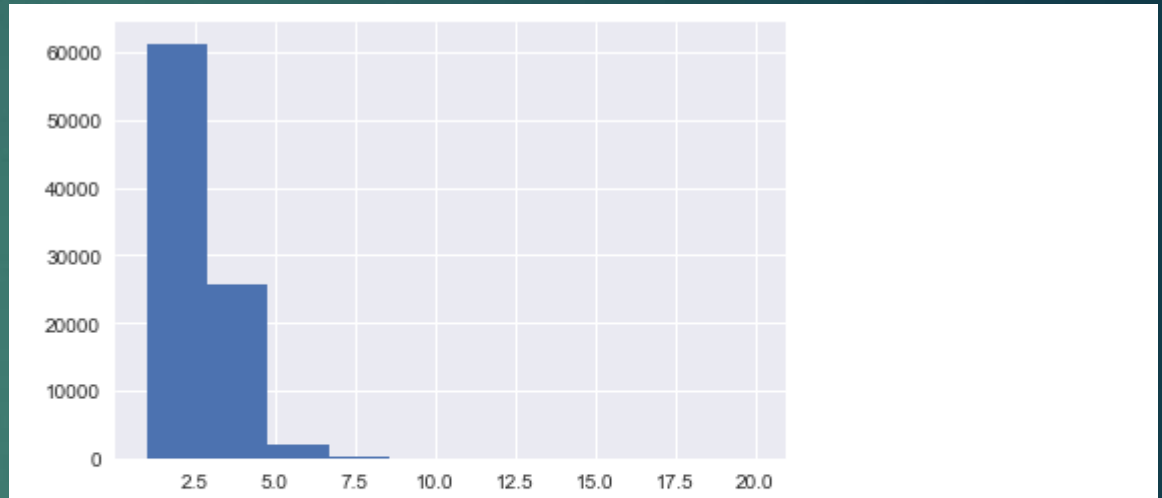
Data Cleaning and Feature Engineering

- *bedroomcnt:*
- *(Number of bedrooms in home)*
- *0.59% NaN*
- *Mean is 3.03, which is not a possible datapoint for bathroom count.*
- *Fill NaN values with median value of 3.0*



Data Cleaning and Feature Engineering

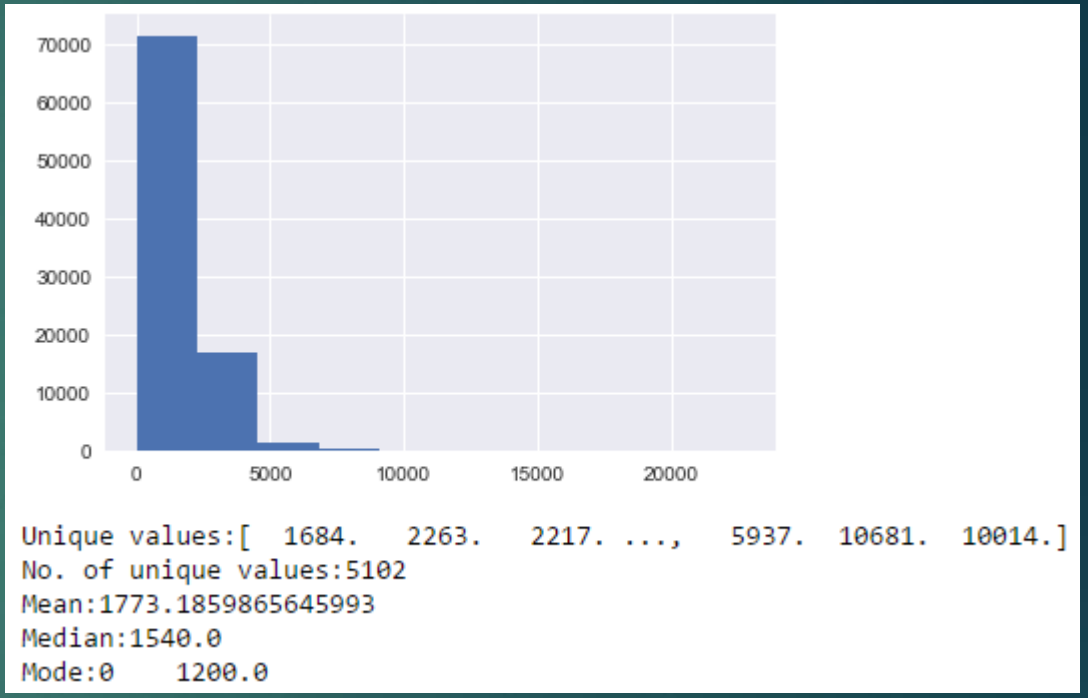
- *Calculatedbathnbr:*
- *(this data is described in data dictionary the same as bathroomcnt, but data is different, keep for now)*
- *0.59% NaN*
- *Mean is 2.309, which is not a possible datapoint for bathroom count.*
- *Fill NaN values with median value of 2.0*




```
[  2.    3.5   3.    2.5   4.    1.    5.    5.5   1.5   8.    nan   4.5
   9.    7.    6.   10.    6.5   7.5  12.   11.   20.    8.5  15. ]
Mean:2.309216212272569
Median:2.0
Mode:0    2.0
dtype: float64
```

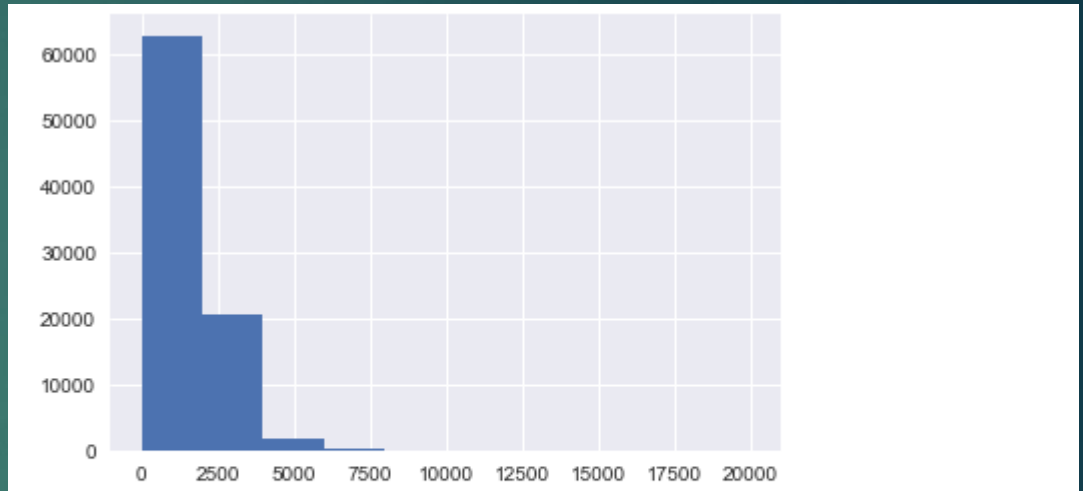
Data Cleaning and Feature Engineering

- *Calculatedfinishedsquarefeet:*
- *(Calculated total finished living area of the home)*
- *1.31% NaN*
- *Mean is 1773, data is continuous*
- *Median is 1540.*
- *Fill NaN values with mean value of 1773 as data is continuous.*



Data Cleaning and Feature Engineering

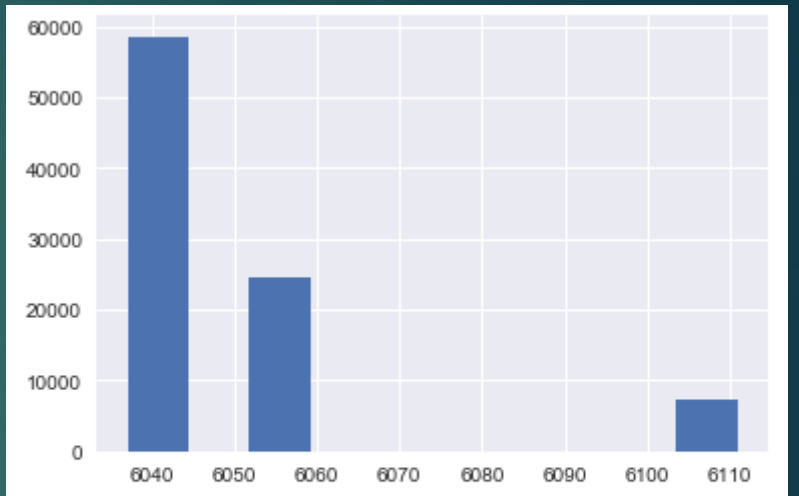
- *Finishedsquarefeet12:*
- *(Finished living area)*
- *5.74% Nan*
- *Mean is 1745*
- *Median is 1518*
- *Fill with the mean value of 1745 as the data is continuous.*




```
Unique values:[  1684.   2263.   2217. ...,   3758.  10681.  10014.]
No. of unique values:4982
Mean:1745.4545305855413
Median:1518.0
Mode:0    1200.0
```

Data Cleaning and Feature Engineering

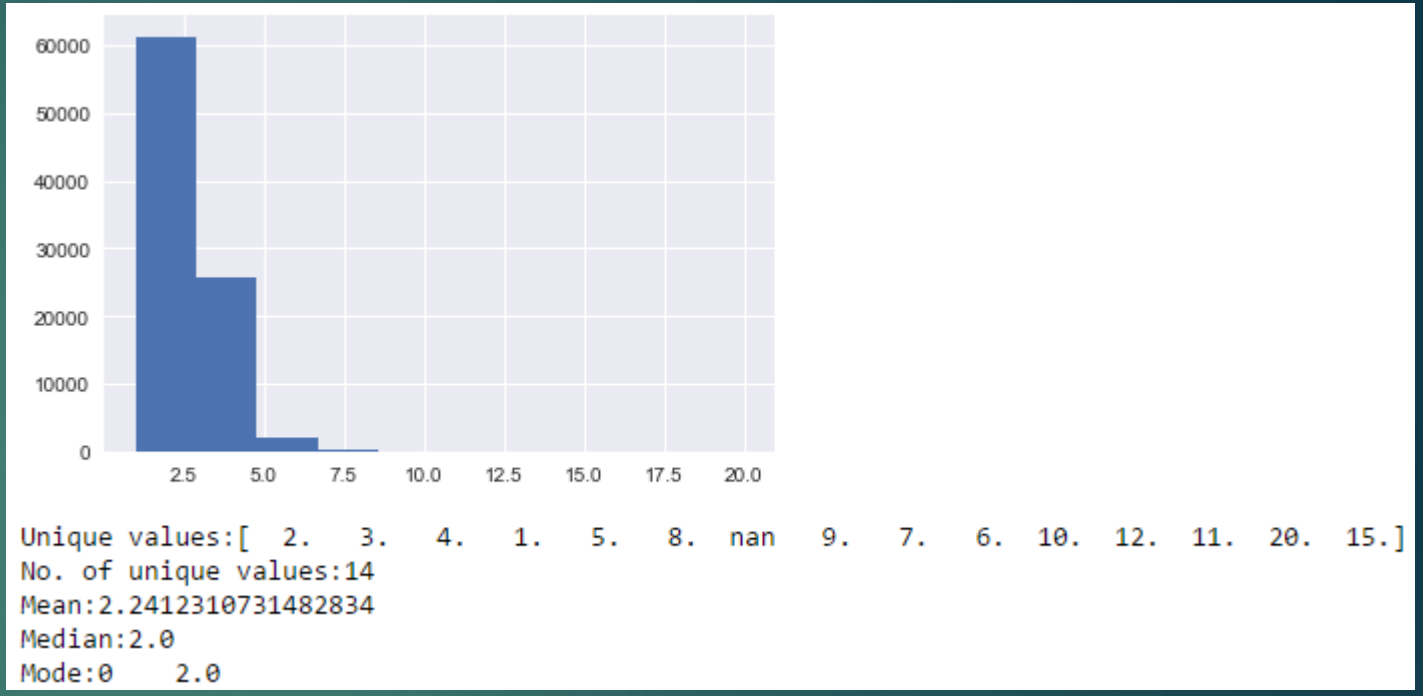
- *Fips*
- *(Federal Information Processing Standard code)*
- *0.59% NaN*
- *Only 3 unique values*
- *Categorical Data similar to zip code.*
- *Can exclude from predictive modeling.*




```
Unique values:[ 6037.  6059.  6111.    nan]
No. of unique values:3
Mean:6048.87055109388
Median:6037.0
Mode:0    6037.0
```

Data Cleaning and Feature Engineering

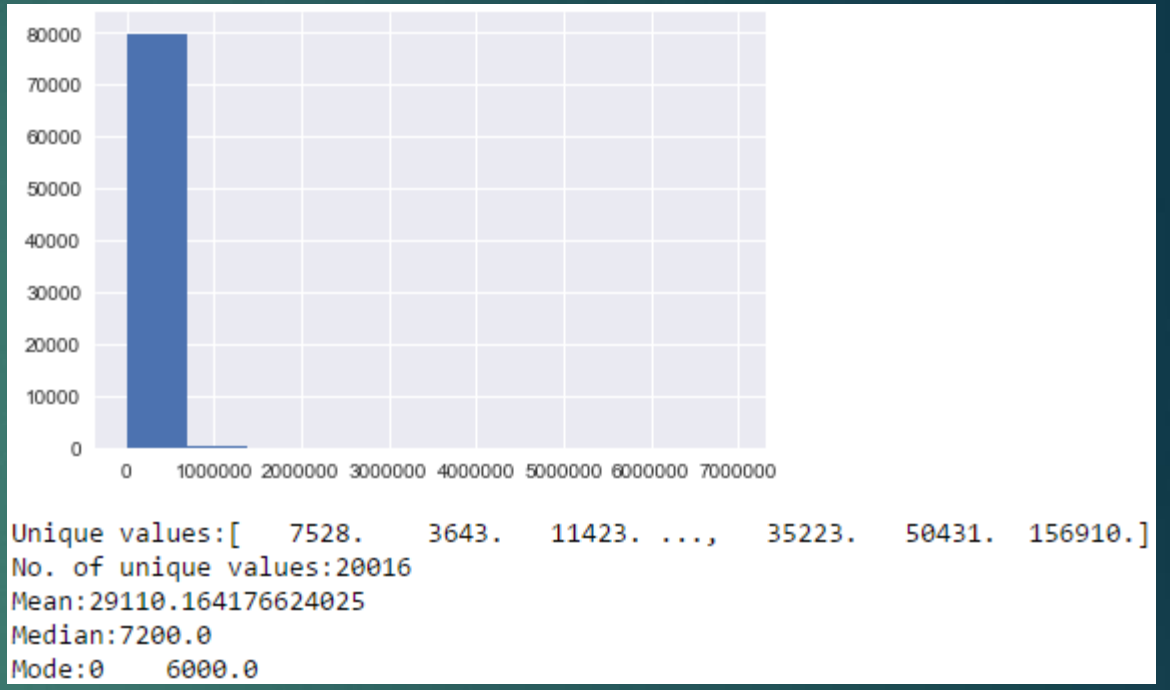
- *Fullbathcnt*
- *(Number of full bathrooms (sink, shower + bathtub, and toilet) present in home)*
- *1.89% NaN*
- *Mean is 2.24*
- *Median is 2*
- *Fill with median to minimize bias towards high end and because 2.24 bathroom cannot exist.*




```
Unique values:[ 2.  3.  4.  1.  5.  8. nan  9.  7.  6. 10. 12. 11. 20. 15.]
No. of unique values:14
Mean:2.2412310731482834
Median:2.0
Mode:0    2.0
```

Data Cleaning and Feature Engineering

- *Lotsizesquarefeet*
- *(Area of the lot in square feet)*
- *Mean is 29,110*
- *Median is 7,200*
- *Fill Nan with median value as mean value would be heavily biased to the high end.*



Data Cleaning and Feature Engineering

- *Propertycountylandusecode*
- *(County land use code i.e. it's zoning at the county level)*
- *77 unique alphanumeric values.*
- *0.591% Nan*
- *Categorical data, can be kept out of modeling.*

```
Unique values:['0100' '1' '010C' '122' '1129' '34' '1128' '010E' '0104' '0101' '0200'
 '0700' '1111' '01DC' '010D' '1110' '0400' '012C' '010V' '1116' '01HC'
 '010G' '0300' '010F' '1117' '0103' '38' '1210' '0111' '010M' '96' '135'
 '0108' '1014' '1112' '0201' '0109' '1310' '010H' '1410' '1222' '1321'
 '1720' '1011' '1432' '0401' '0102' '012D' '73' '105' '0110' '100V' '0130'
 '8800' '0303' '0210' '1012' '1333' '0114' '01DD' '020G' '040A' '012E'
 '020M' '040V' '070D' '1200' '030G' '1722' '6050' '1421' '010' nan '200'
 '0' '1420' '0131' '0301']
No. of unique values:77
```

Data Cleaning and Feature Engineering

- *regionidcity*
- *(City in which the property is located (if any))*
- *Categorical data, can be excluded from modeling.*

```
Unique values:[  12447.   47019.   17686.   29712.   24174.   13150.   25459.   46098.
  396054.   52650.   17150.   25218.   53655.   46298.   34780.   47568.
   24832.    5465.   40227.   54311.   10389.   21412.    6395.   33252.
   25458.   24384.   20008.   33836.    8384.   24812.   53571.   51617.
   32380.   45888.   45457.   15554.   24245.   16764.   27110.   40081.
   41673.   34278.   12773.   16389.   42150.   54970.   52842.   34543.
   15237.   53636.   37688.   13693.    5534.   54722.   50749.   27491.
  118225.   27103.   13091.   33837.   50677.   10608.   10723.   48424.
   47762.    6021.    9840.   18874.   38032.   44833.   24435.   12292.
   10774.  396556.   45602.   33311.   33612.   44116.   10241.   25974.
   21778.   14634.   11626.   40009.   30187.   32923.   26483.   26531.
   14906.   14111.   26964.   18875.   30908.   13716.   39306.  118914.
   38980.   25621.   51861.  118878.   34636.    4406.   51239.   17882.
   30399.   37015.   46314.   29189.   12520.   14542.   19177.   54053.
   37086.   54299.  396053.   52835.   53027.   39308.   47547.  116042.
  118694.   32616.   42967.   16677.  118994.   22827.   33840.  396550.
  113576.   34037.  118875.   45398.  118217.   54212.  114828.  396551.
   40110.   46178.   47695.   27183.   55753.   25953.   10734.  118895.
   46080.  272578.   47198.   17597.   13232.    3491.   39076.   33727.
   30267.   32753.   56780.   54352.   42091.  114834.   26965.   32927.
   25468.   16961.  113412.   36502.   18098.   13311.   33312.   25271.
    6822.   10815.   53162.   31134.  118880.   37882.   24797.   21395.
    6285.]
No. of unique values:177
Mean:33353.366425409
Median:24832.0
Mode:0    12447.0
```

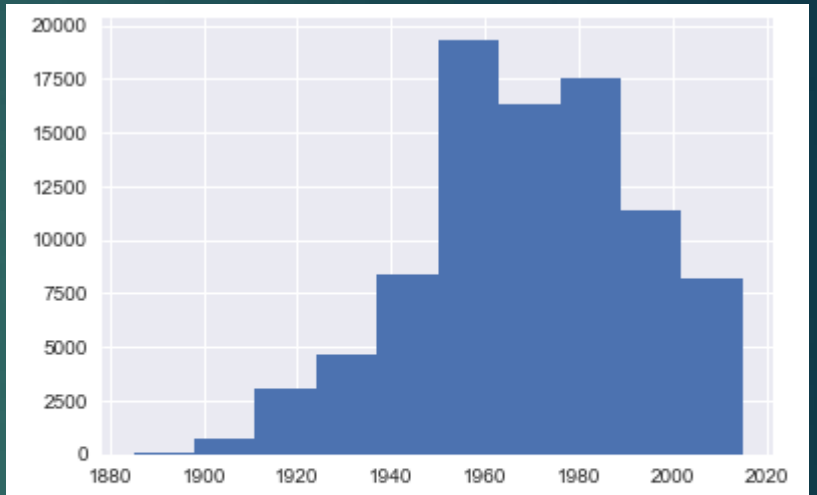
Data Cleaning and Feature Engineering

- *regionidzip*
- *Zip code in which the property is located*
- *Categorical data, can be excluded from model*

```
No. of unique values:388
Mean:96585.86597374789
Median:96393.0
Mode:0    97319.0
```

Data Cleaning and Feature Engineering

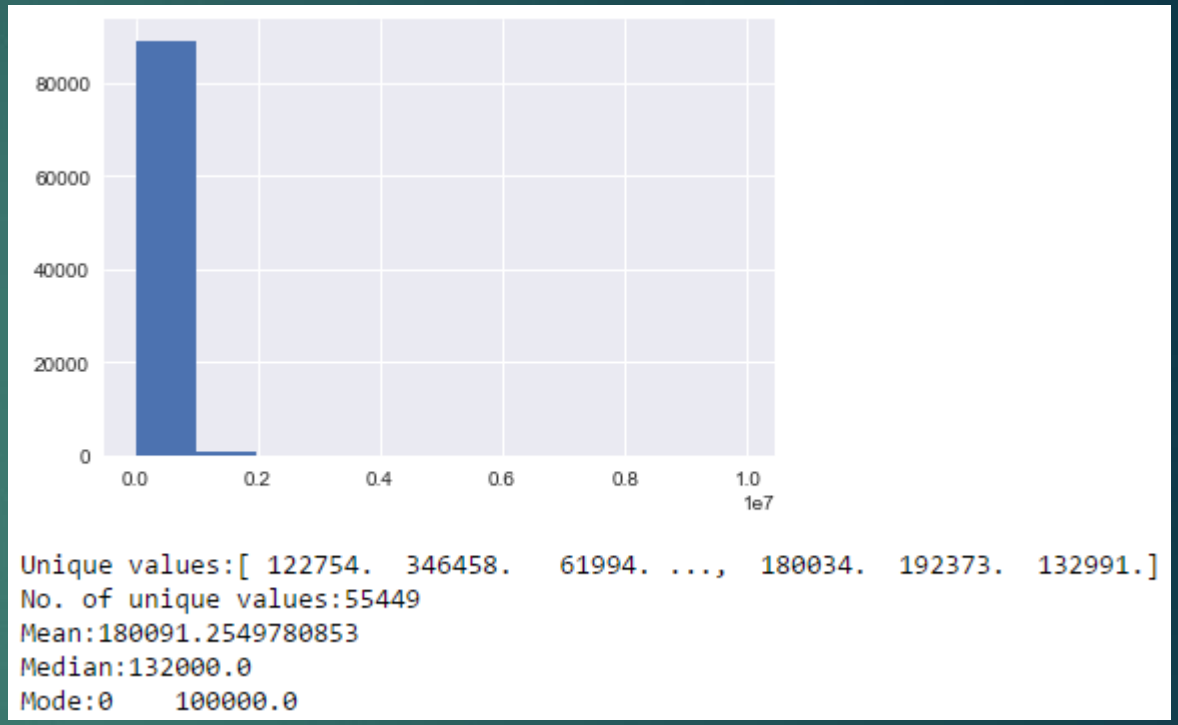
- *Yearbuilt*
- *(The Year the principal residence was built)*
- *1.42% Nan*
- *Mean is 1968.45*
- *Median is 1970*
- *Fill with the median as the rest of the data is interval data as well.*




```
No. of unique values:130
Mean:1968.4501221157814
Median:1970.0
Mode:0    1955.0
```

Data Cleaning and Feature Engineering

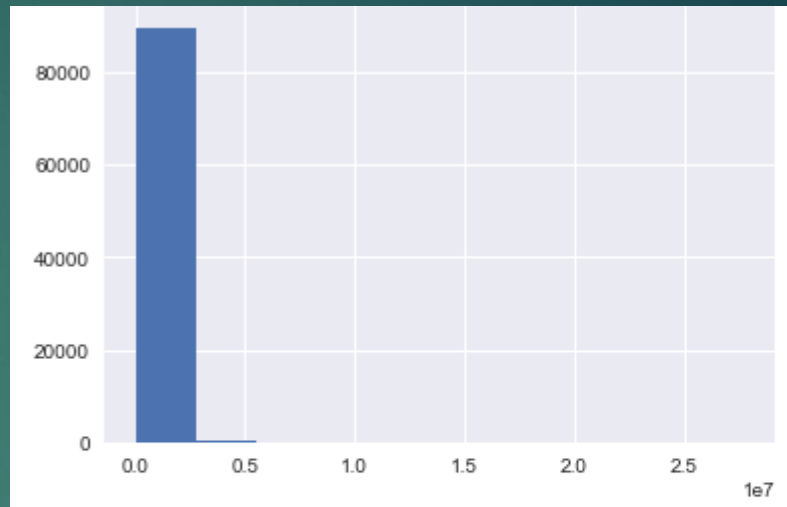
- *Structuretaxvaluedollarcnt*
- *(The assessed value of the built structure on the parcel)*
- *1.008% Nan*
- *Mean is 180,091*
- *Median is 132,000*
- *Fill Nan values with median value of 132,000 to prevent bias towards small group of very high priced properties.*




```
Unique values:[ 122754.  346458.   61994. ...,  180034.  192373.  132991.]
No. of unique values:55449
Mean:180091.2549780853
Median:132000.0
Mode:0    100000.0
```

Data Cleaning and Feature Engineering

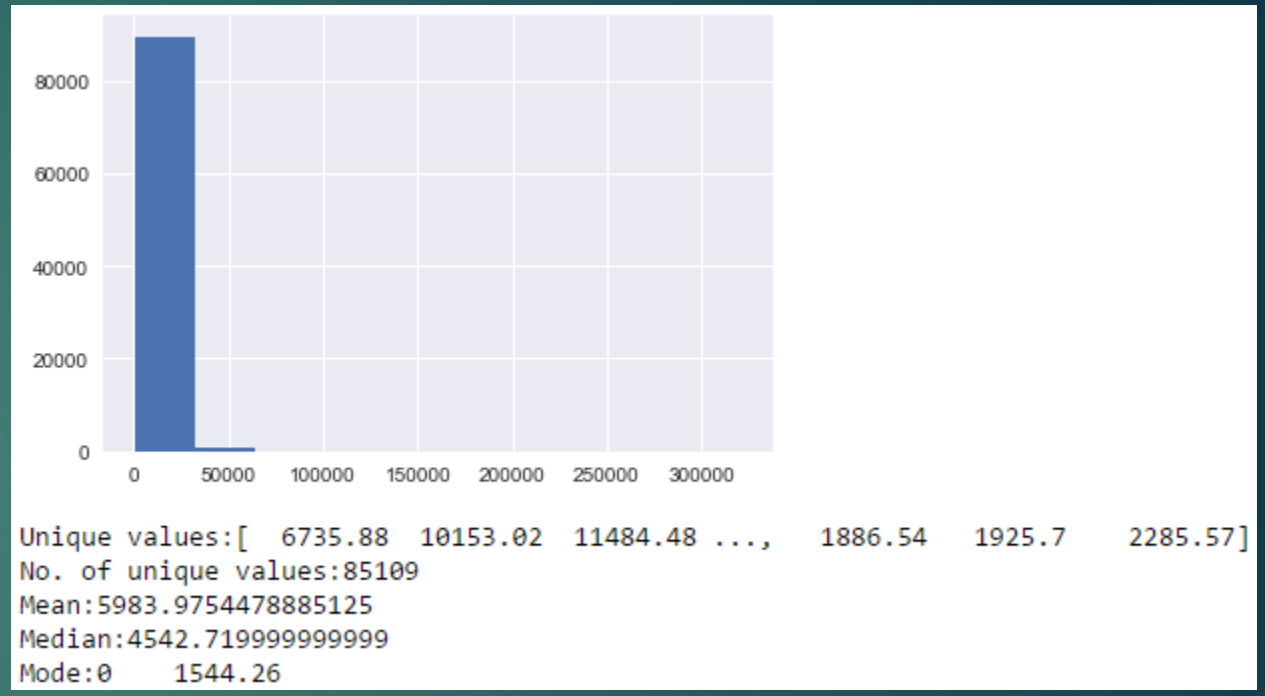
- *Taxvaluedollarcnt*
- *(The total tax assessed value of the parcel)*
- *0.59% NaN*
- *Mean is 457,672*
- *Median is 342,872*
- *Fill with median value of 342,872 to avoid bias towards small group of high tax value properties.*




```
Unique values:[ 360170.  585529.  119906. ...,  315288.   38096.  165869.]
No. of unique values:55937
Mean:457672.44256865286
Median:342872.0
Mode:0    350000.0
1    420000.0
2    500000.0
```

Data Cleaning and Feature Engineering

- *Taxamount*
- *(The total property tax assessed for that assessment year)*
- *0.596% Nan*
- *Mean is 5,983*
- *Median is 4,542*
- *Fill Nan with median to prevent bias towards small group of expensive properties*

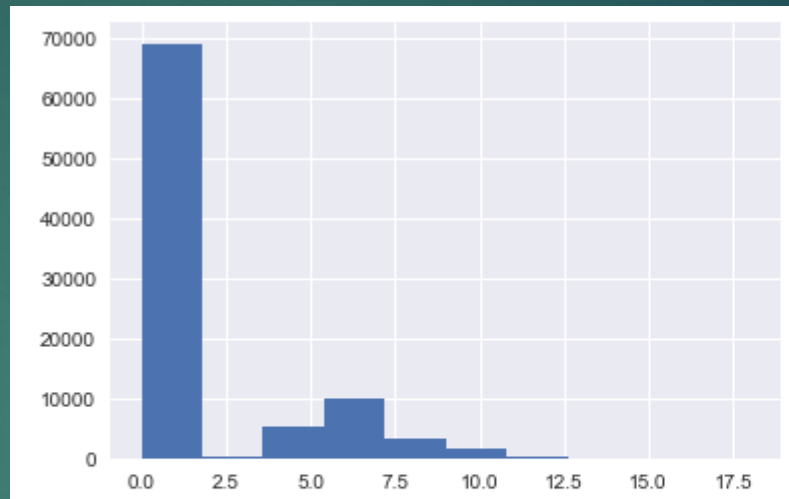


Data Cleaning and Feature Engineering

- *Censustractandblock*
- *(Census tract and block ID combined - also contains blockgroup assignment by extension)*
- *Categorical data, can be excluded from modeling.*

Data Cleaning and Feature Engineering

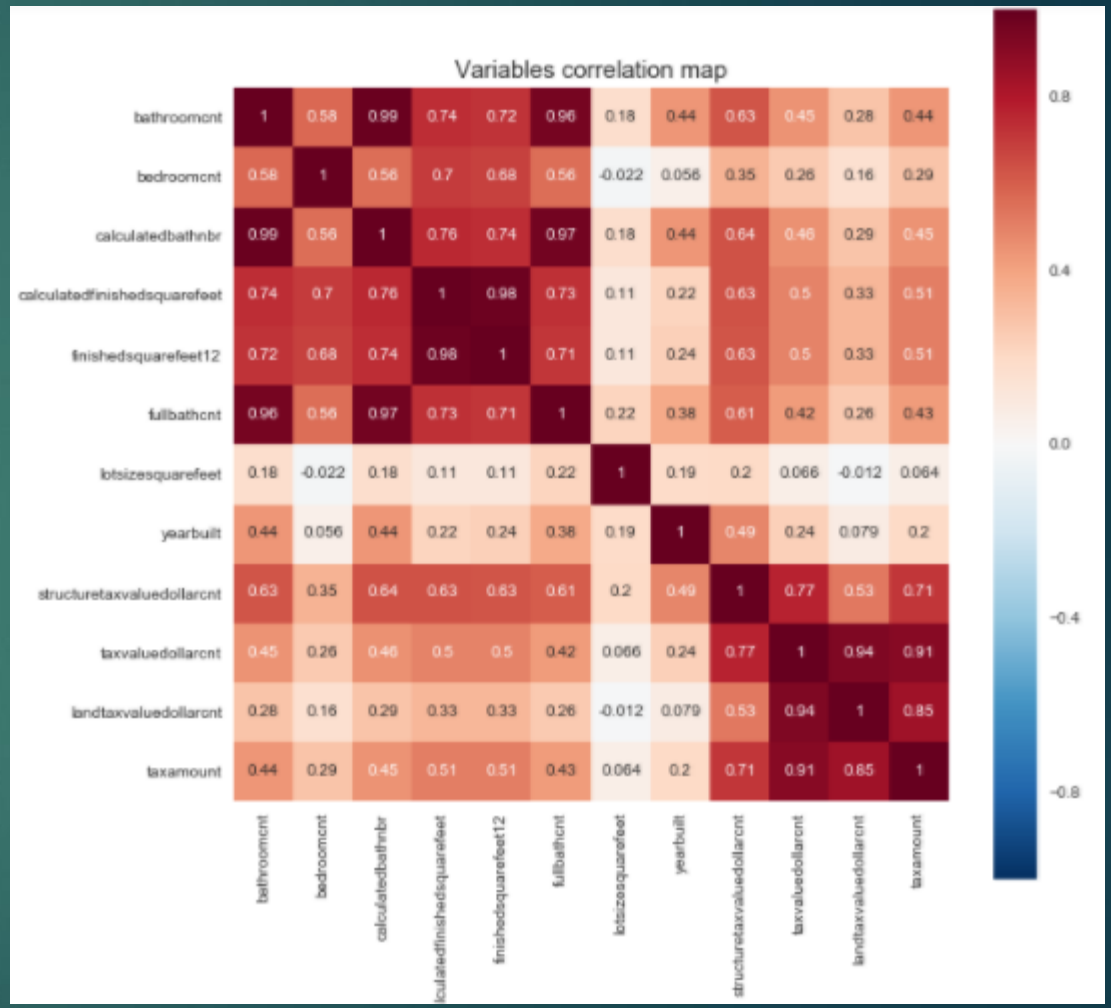
- *Roomcnt*
- *(Total number of rooms in the principal residence)*
- *The median is zero and the mean is 1.48. Data makes no sense, not possible for median to be zero when value stands for number of rooms in a building. Data is useless. Not worth keeping.*




```
Unique values:[  0.   8.   6.   5.   7.   4.   3.   9.  12.  11.  10.   2.   1.  13.  15.
  14.  18.]
No. of unique values:17
Mean:1.4877828457995517
Median:0.0
Mode:0    0.0
```

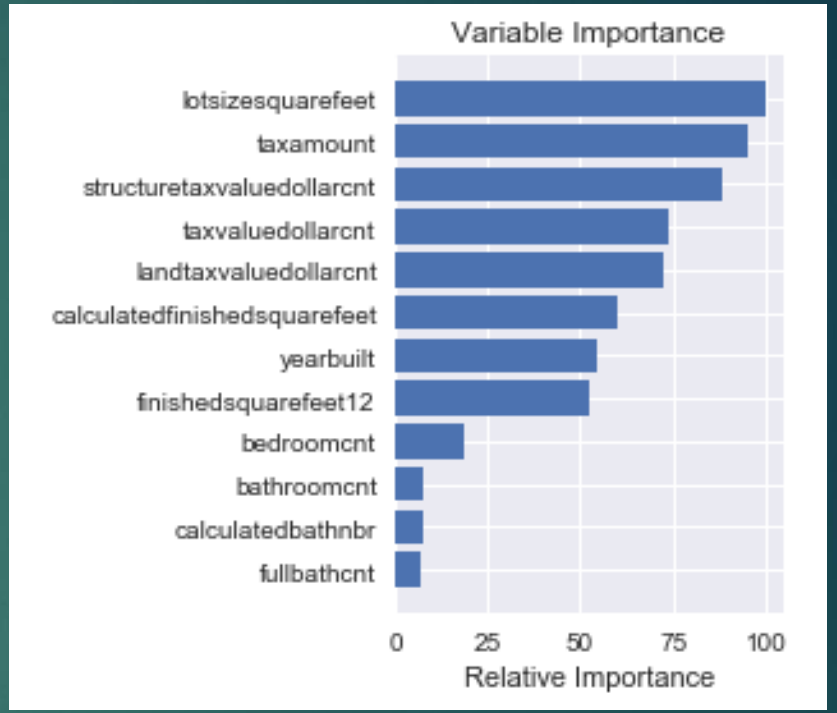
Correlation Analysis

- *Heat map shows heavily correlation in 2 blocks:*
- *bedroomcnt, bathroomcnt, calculatedbathbr, calculatedfinishedsquarefeet, fullbathcnt are all highly correlated.*
- *Structuretaxdollarvaluecnt, taxvaluedollarcnt, landtax valuedollarcnt, and tax amount all highly correlated.*
- *Yearbuilt and lotsizesquarefeet stand alone*
- *Potentially narrows data down to 4 features.*



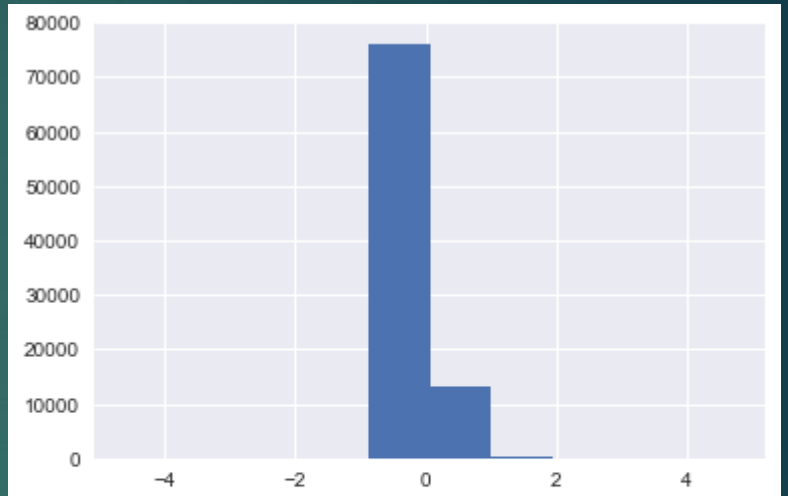
Variable Importance

- *Running a quick random forests and extracting variable importance to prediction of log error shows that the highly correlated group that relates to taxes are all very influential.*
- *Our stand alone variables of lot size square feet and year built also seem to hold some weight.*
- *The correlated group that related to structure size such as bedroom and bathroom count don't seem to show strong influence.*



Target Data Analysis

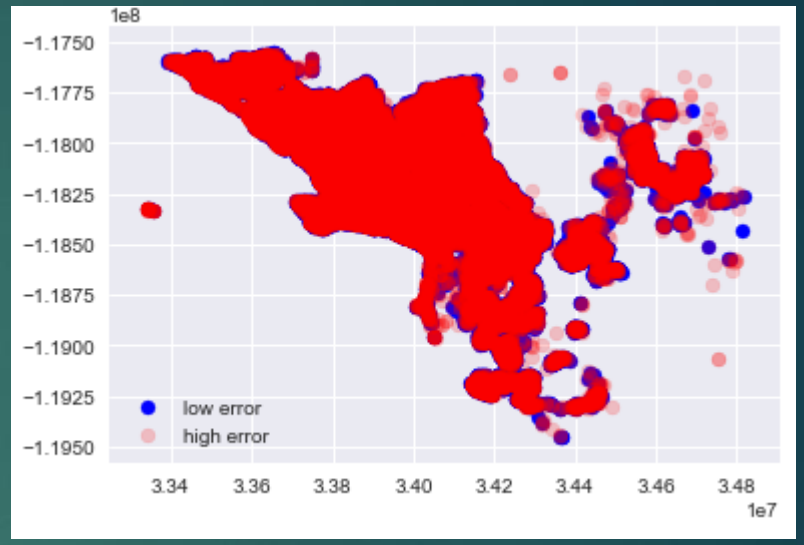
- *Logerror*
- *While there is a range of -4.6 to 4.6 more than half the data falls between -0.025 and 0.039. That's a very tight range for most of the data to fall in.*

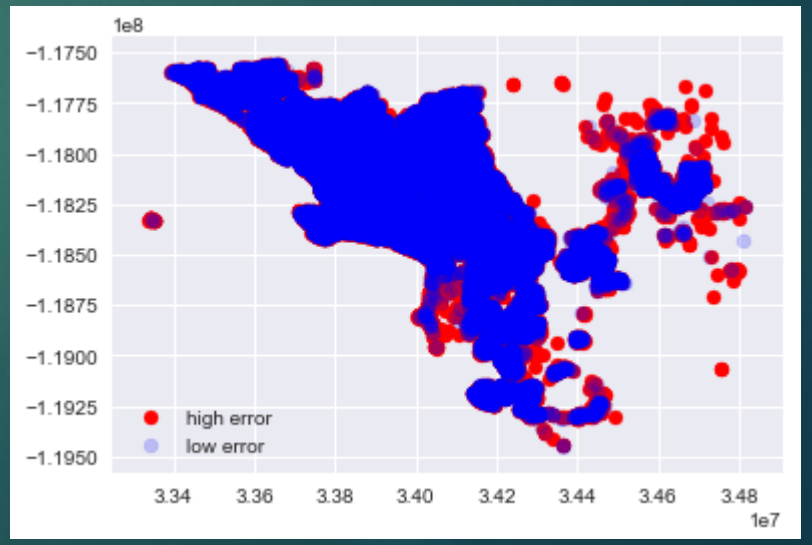


count	89669.000000
mean	0.011682
std	0.160438
min	-4.605000
25%	-0.025300
50%	0.006000
75%	0.039200
max	4.737000

Target Data Analysis -Latitude and Longitude

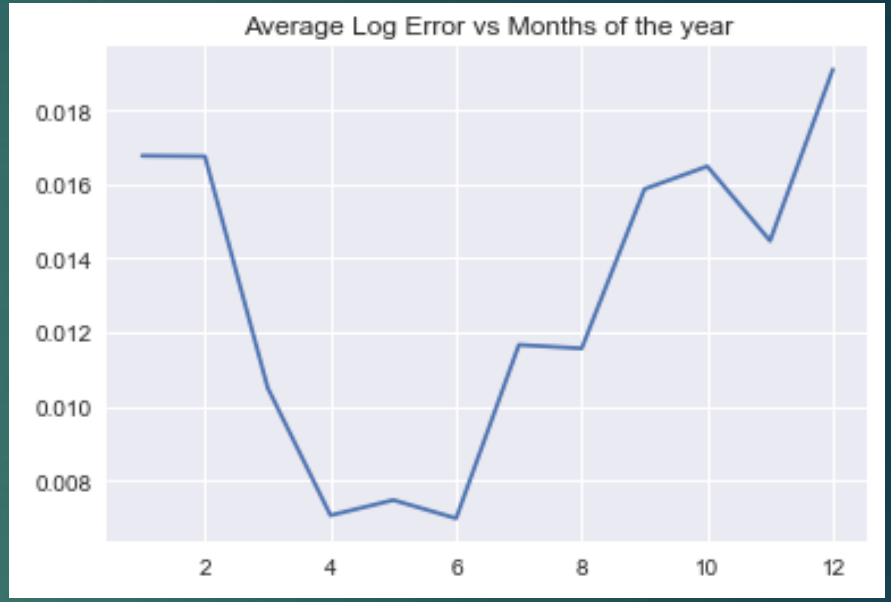
- *Low Error defined as between -0.0253 and 0.0392.*
- *High error defined as the remaining data.*
- *No clear geographical pattern for high vs low error predictions.*





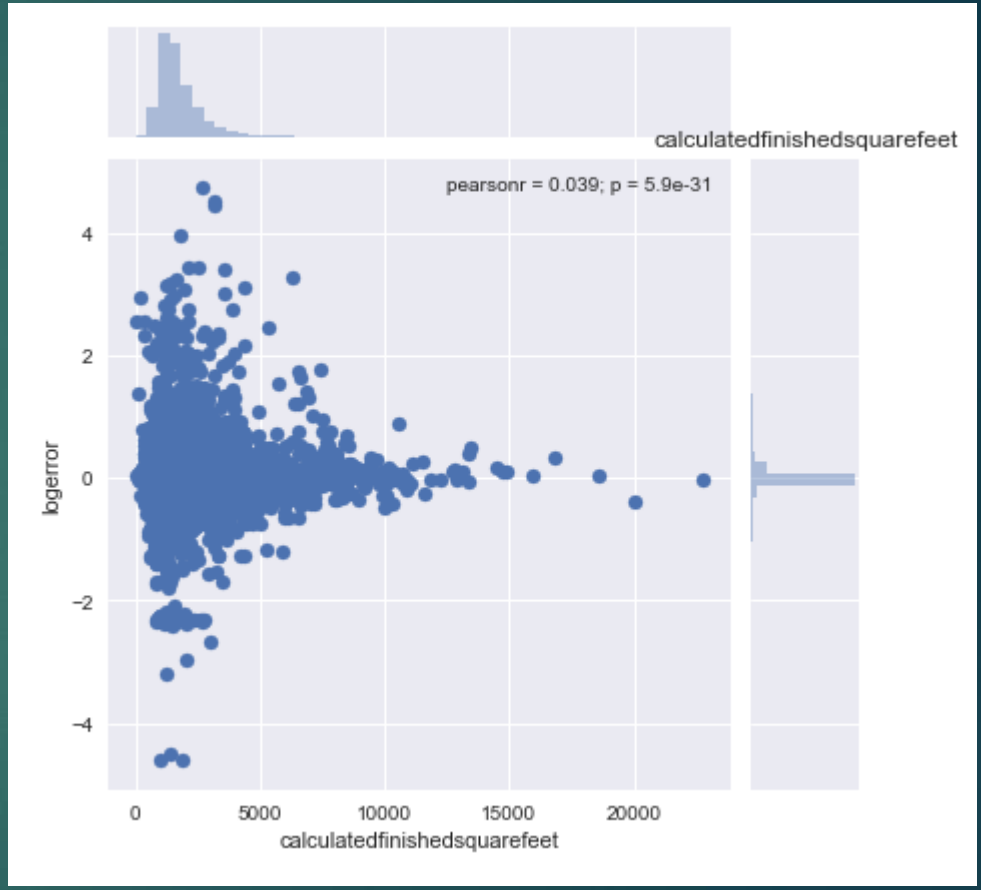
Target Data Analysis -Month of sale

- *The average log error in the winter can be twice as high as in the summer.*



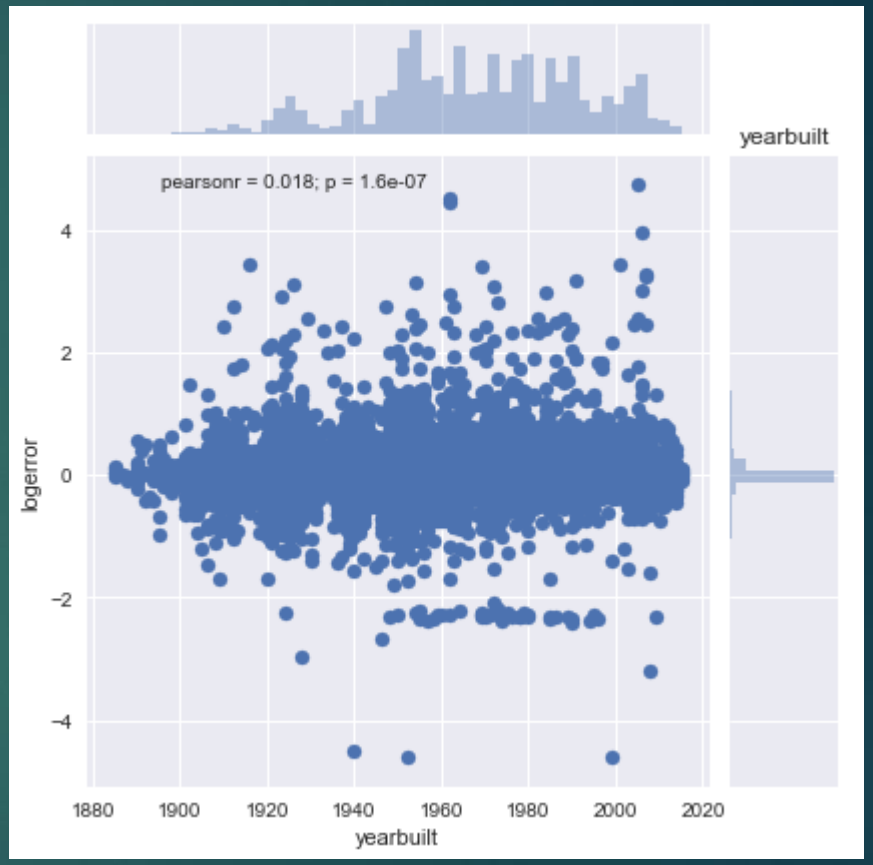
Examination of Influential Features compared to logerror

- *Calculatedfinishedsquarefeet*
- *Smaller properties seems to account for nearly all of the log error. However the vast majority of properties sold are small properties.*



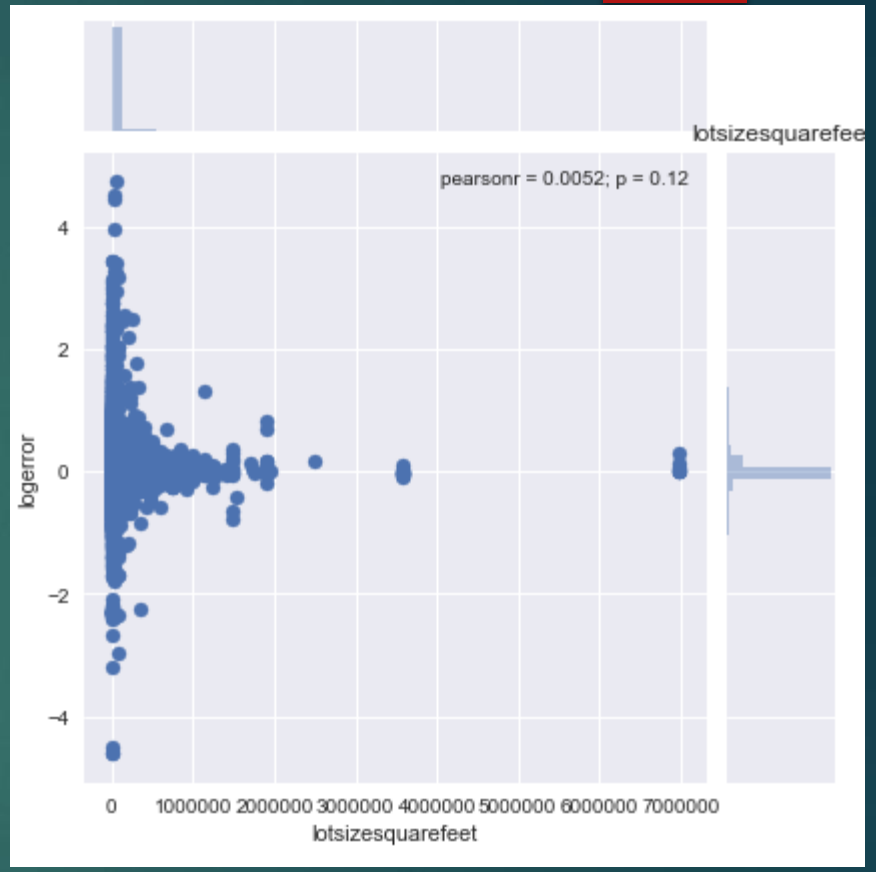
Examination of Influential Features compared to logerror

- *Yearbuilt*
- *Log error seems pretty evenly dispersed along year built.*



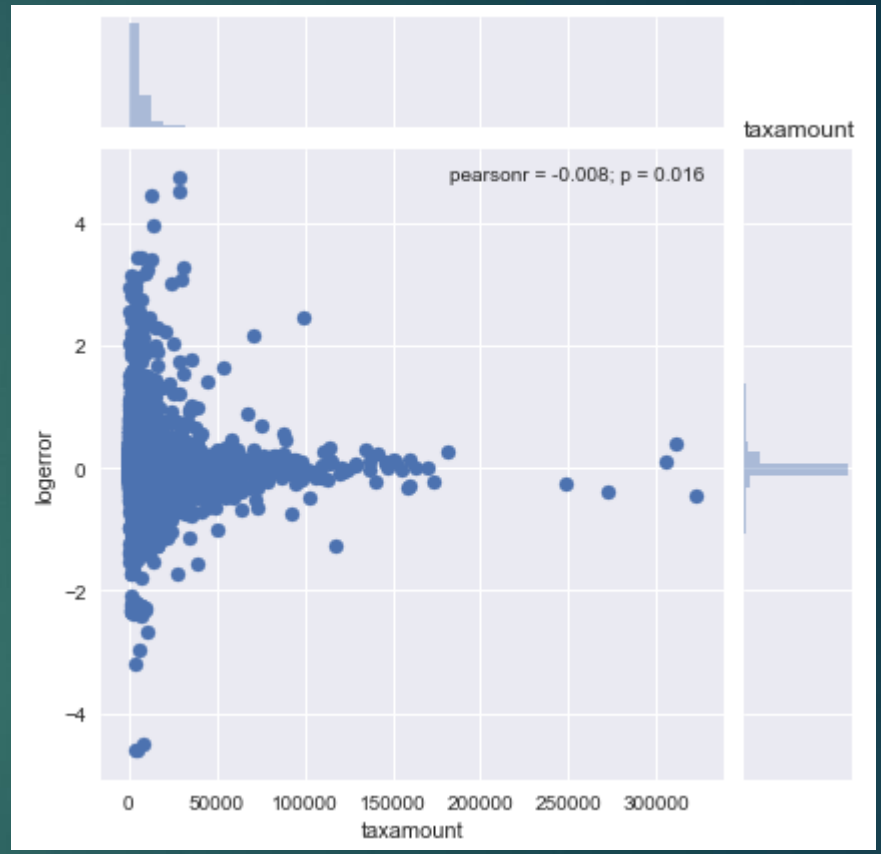
Examination of Influential Features compared to logerror

- *Lotsizesquarefeet*
- *Similar to calculatedfinishedsquarefeet, smaller properties showing greater range in log error.*



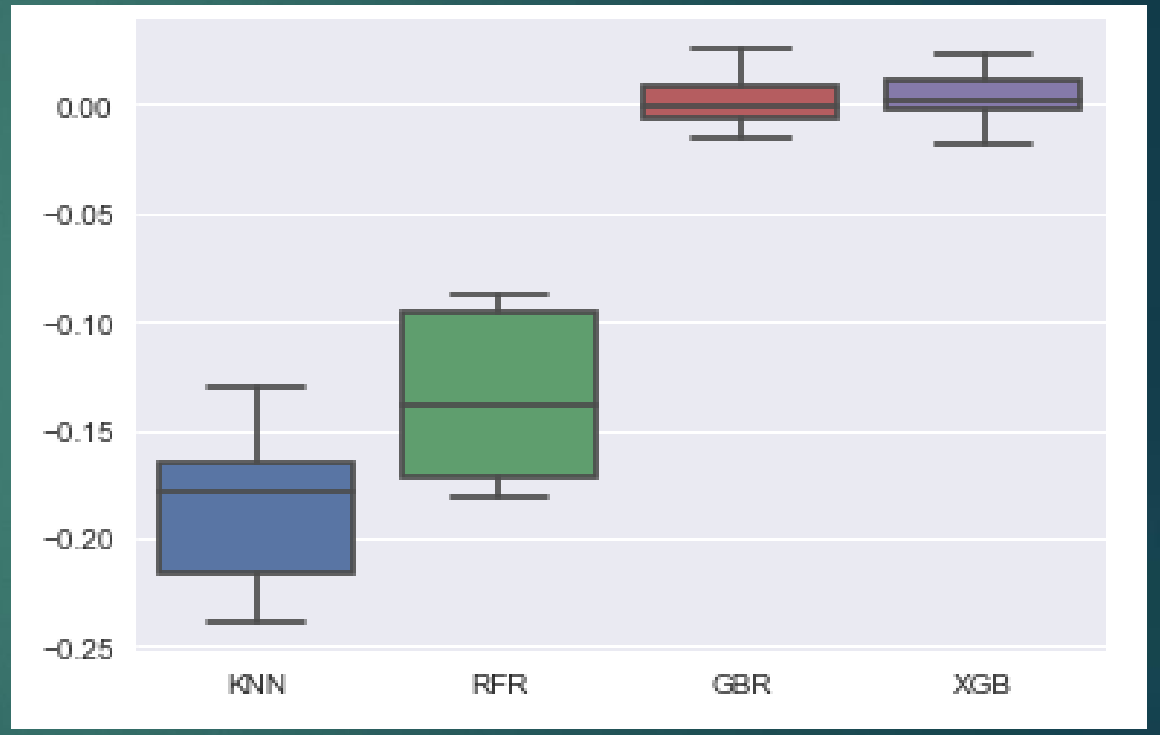
Examination of Influential Features compared to logerror

- *Taxamount*
- *Similar to calculatedfinishedsquarefeet, smaller properties showing greater range in log error.*



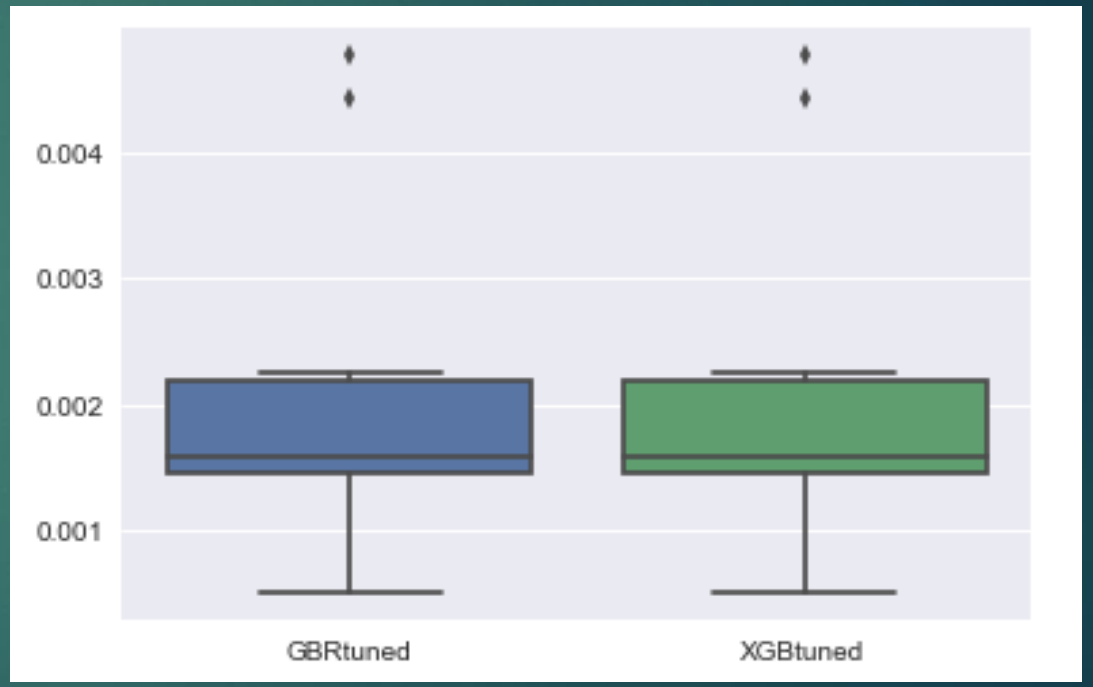
Model Building

- *Cut down to 4 features:*
  - *Tax Amount*
  - *Lot Size Square Feet*
  - *Year built*
  - *Calculated Finished Square Feet*
- *Test KNN, RFR, Gradient Boosting and XGBoost with 10 folds*



Hyper parameter Tuning

- *Focus on Gradient Boosting and XGBoost*



Conclusion

- *Hyper parameter tuning on gradient and xgboost provided limited to no improvements, primarily resulting in a decrease variance in cross fold validation R-squared values.*
- *Optimal hyper parameter tuning results in only modest R-squared values, barely above zero. Submission of test predictions to competition scores a MAE of 0.065018 vs a baseline of 0.066301.*
- *Zillow's Zestimates seem to be very accurate already, and further improvements of significant value seem difficult if not impossible to achieve.*

Future Improvements

- *Further gains could be achieved potentially by lowering the threshold on NaN feature sets.*
- *Further experimentation on relation of logerror as it relates to month of sale could result in further improvement.*