test

edris safari

2/25/2020

# Final Project

The purpose of this project is to use techniques learned in this class to exercise exploratory data analysis on a given data set, The dataset chosen for this project are:

* properties\_2016.csv
* transactions\_2016.csv

properties dataset contins the information about individual homes that were sold in 2016. The transaction dataset has the transaction date when the house was sold and the log error from the sales price estimated by zillow(zestimate).

## Summary of transactions data set

## parcelid logerror date   
## Min. : 10711738 Min. :-4.60500 2016-07-29: 910   
## 1st Qu.: 11559500 1st Qu.:-0.02530 2016-04-29: 902   
## Median : 12547337 Median : 0.00600 2016-09-30: 894   
## Mean : 12984656 Mean : 0.01146 2016-06-30: 874   
## 3rd Qu.: 14227552 3rd Qu.: 0.03920 2016-05-27: 863   
## Max. :162960842 Max. : 4.73700 2016-08-31: 737   
## (Other) :85095

## head of transactions data set

## parcelid logerror date  
## 1 11016594 0.0276 2016-01-01  
## 2 14366692 -0.1684 2016-01-01  
## 3 12098116 -0.0040 2016-01-01  
## 4 12643413 0.0218 2016-01-02  
## 5 14432541 -0.0050 2016-01-02  
## 6 11509835 -0.2705 2016-01-02

## head of properties data set

## X parcelid aircon architectural\_style area\_basement num\_bathroom  
## 1 1234018 11073844 1 NA NA 3  
## 2 2580667 12681060 NA NA NA 2  
## 3 884268 12298768 NA NA NA 2  
## 4 1072195 10836492 NA NA NA 2  
## 5 609104 14126112 NA NA NA 2  
## 6 972322 11973991 NA NA NA 1  
## num\_bedroom framing quality num\_bathroom\_calc deck area\_firstfloor\_finished  
## 1 5 NA 4 3 NA NA  
## 2 3 NA 7 2 NA NA  
## 3 3 NA 7 2 NA NA  
## 4 4 NA 7 2 NA NA  
## 5 2 NA NA 2 NA NA  
## 6 2 NA 7 1 NA NA  
## area\_total\_calc area\_live\_finished area\_liveperi\_finished area\_total\_finished  
## 1 3036 NA NA 3036  
## 2 1158 1158 NA NA  
## 3 1348 1348 NA NA  
## 4 1610 NA NA 1610  
## 5 1212 1212 NA NA  
## 6 979 979 NA NA  
## area\_unknown area\_base fips num\_fireplace num\_bath num\_garage area\_garage  
## 1 NA NA 6037 NA 3 NA NA  
## 2 NA NA 6037 NA 2 NA NA  
## 3 NA NA 6037 NA 2 NA NA  
## 4 NA NA 6037 NA 2 NA NA  
## 5 NA NA 6059 NA 2 2 400  
## 6 NA NA 6037 NA 1 NA NA  
## flag\_tub heating latitude longitude area\_lot num\_pool area\_pool pooltypeid10  
## 1 NA 34266433 -118515891 NA NA NA NA  
## 2 2 33817148 -118366118 6355 NA NA NA  
## 3 7 33962939 -118226118 5094 NA NA NA  
## 4 NA 34192310 -118506742 8099 NA NA NA  
## 5 NA 33848859 -118024802 6000 NA NA NA  
## 6 7 34123857 -118201009 5200 1 NA NA  
## pooltypeid2 pooltypeid7 zoning\_landuse\_county zoning\_landuse zoning\_property  
## 1 NA NA 0200 246 LAR1  
## 2 NA NA 0100 261 TORR-LO  
## 3 NA NA 0100 261 LCR305  
## 4 NA NA 0200 246 LAR2  
## 5 NA NA 122 261   
## 6 NA 1 0101 261 LAR1  
## rawcensustractandblock region\_city region\_county region\_neighbor region\_zip  
## 1 60371112 12447 3101 31817 96370  
## 2 60376512 54722 3101 416967 96163  
## 3 60375348 39076 3101 NA 96110  
## 4 60371320 12447 3101 32368 96415  
## 5 60591103 10608 1286 NA 96180  
## 6 60371833 12447 3101 115609 96023  
## num\_room story num\_75\_bath material num\_unit area\_patio area\_shed build\_year  
## 1 0 NA NA NA 2 NA NA 2008  
## 2 0 NA NA NA 1 NA NA 1956  
## 3 0 NA NA NA 1 NA NA 1925  
## 4 0 NA NA NA 2 NA NA 1948  
## 5 6 NA NA NA NA NA NA 1955  
## 6 0 NA NA NA 1 NA NA 1941  
## num\_story flag\_fireplace tax\_building tax\_total tax\_year tax\_land  
## 1 NA 361423 411051 2015 49628  
## 2 NA 121581 331215 2015 209634  
## 3 NA 102501 263854 2015 161353  
## 4 NA 54376 144898 2015 90522  
## 5 1 58199 230925 2015 172726  
## 6 NA 61066 264501 2015 203435  
## tax\_property tax\_delinquency tax\_delinquency\_year censustractandblock  
## 1 5234.38 NA 6.037111e+13  
## 2 4017.45 NA 6.037651e+13  
## 3 5107.38 NA 6.037535e+13  
## 4 1993.37 NA 6.037132e+13  
## 5 2851.24 NA 6.059110e+13  
## 6 3231.99 NA 6.037183e+13

## Column names

### properties

X, parcelid, aircon, architectural\_style, area\_basement, num\_bathroom, num\_bedroom, framing, quality, num\_bathroom\_calc, deck, area\_firstfloor\_finished, area\_total\_calc, area\_live\_finished, area\_liveperi\_finished, area\_total\_finished, area\_unknown, area\_base, fips, num\_fireplace, num\_bath, num\_garage, area\_garage, flag\_tub, heating, latitude, longitude, area\_lot, num\_pool, area\_pool, pooltypeid10, pooltypeid2, pooltypeid7, zoning\_landuse\_county, zoning\_landuse, zoning\_property, rawcensustractandblock, region\_city, region\_county, region\_neighbor, region\_zip, num\_room, story, num\_75\_bath, material, num\_unit, area\_patio, area\_shed, build\_year, num\_story, flag\_fireplace, tax\_building, tax\_total, tax\_year, tax\_land, tax\_property, tax\_delinquency, tax\_delinquency\_year, censustractandblock

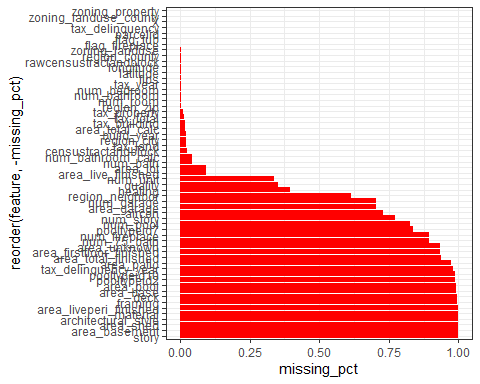
### transactions

parcelid, logerror, date

## statistics of logerror in transactions

* Mean : 0.0114572
* median : 0.006
* std : ’ 0.1610788

## Features with percentage of missing values



## Selected features

X, parcelid, aircon, num\_bathroom, num\_bedroom, quality, num\_bathroom\_calc, area\_total\_calc, area\_live\_finished, fips, num\_bath, num\_garage, area\_garage, flag\_tub, heating, latitude, longitude, area\_lot, zoning\_landuse\_county, zoning\_landuse, zoning\_property, rawcensustractandblock, region\_city, region\_county, region\_neighbor, region\_zip, num\_room, num\_unit, build\_year, flag\_fireplace, tax\_building, tax\_total, tax\_year, tax\_land, tax\_property, tax\_delinquency, censustractandblock

