Zillow Prize: Zillow's Home Value Prediction (Zestimate)

A CASE STUDY REPORT

Mayank Aggarwal

Mayank953@gmail.com
Github

TABLE OF CONTENTS

ABOUT ZILLOW

Zillow is an online real estate database company founded in 2006 – Wikipedia

Zillow is the leading real estate and rental marketplace dedicated to empowering consumers with data, inspiration and knowledge around the place they call home, and connecting them with the best local professionals who can help. – Zillow.com

Zestimate:

"Zestimates" are estimated home values based on 7.5 million statistical and machine learning models that analyze hundreds of data points on each property. And, by continually improving the median margin of error (from 14% at the onset to 5% today),

1. Problem Statement

Our task in this competition is to predict the the difference between the actual price and the estimate of the price (Zestimate). So, in fact we are predicting, where Zillow's Zestimate will be good, and where it will be bad.

However, we don't have to predict a single value, but instead for 6 different time points (from October 2016 to December 2017)

2. INTRODUCTION TO PROJECT

A) CASE STUDY EXPLANATION

Zillow's Zestimate home valuation has shaken up the U.S. real estate industry since first released 11 years ago. The Zestimate was created to give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of home value information at no cost.

In this case study we are required to develop an algorithm that makes predictions about the future sale prices of homes. We will be building a model to improve the Zestimate residual error.

B) DOMAIN KNOWLEDGE

A home is often the largest and most expensive purchase a person makes in his or her lifetime. Ensuring homeowners have a trusted way to monitor this asset is incredibly important.

The home or any property is at a higher level categorized not only on the basis of location or on its area but certain other factor i.e. tax–related, room count , year built, facilities provided. Proper Knowledge & analysis of how these feature affect the price of the house individually or with each other is required. Also computing additional insights/feature from these basic one which can affect the Zestimate is required.

C) EVALUATION

Results are evaluated on Mean Absolute Error between the predicted log error and the actual log error. The log error is defined as

$$\text{logerror}=\log(\text{Zestimate})-\log(\text{SalePrice})$$

it is recorded in the transactions training data. If a transaction didn't happen for a property during that period of time, that row is ignored and not counted in the calculation of MAE.

We are to predict 6 time points for all properties: October 2016 (201610), November 2016 (201611), December 2016 (201612), October 2017 (201710), November 2017 (201711), and December 2017 (201712).

D) DATASET PROVIDED

We are provided with the following data:-

- properties_2016.csv - all the properties with their home features for 2016.
- train_2016.csv - the training set with transactions from 1/1/2016 to 12/31/2016
- zillow_data_dictionary.xlsx – it explains the various features present in the properties dataset while also clears the meaning of numbers use for categorical variable. The meaning of every knowledge can be obtained from here.

3. EXPLORATORY DATA ANALYSIS & DATA EXPLORATION

We are given 2 dataset mainly. Properties (i.e properties_2016) has dimensions of (2985217,58) while the transaction (i.e. train_2016) has the dimensions of (90811,3). The transaction data has target variable i.e logerror which is associated with each & every parcelid.

PROPERTIES DATASET:-

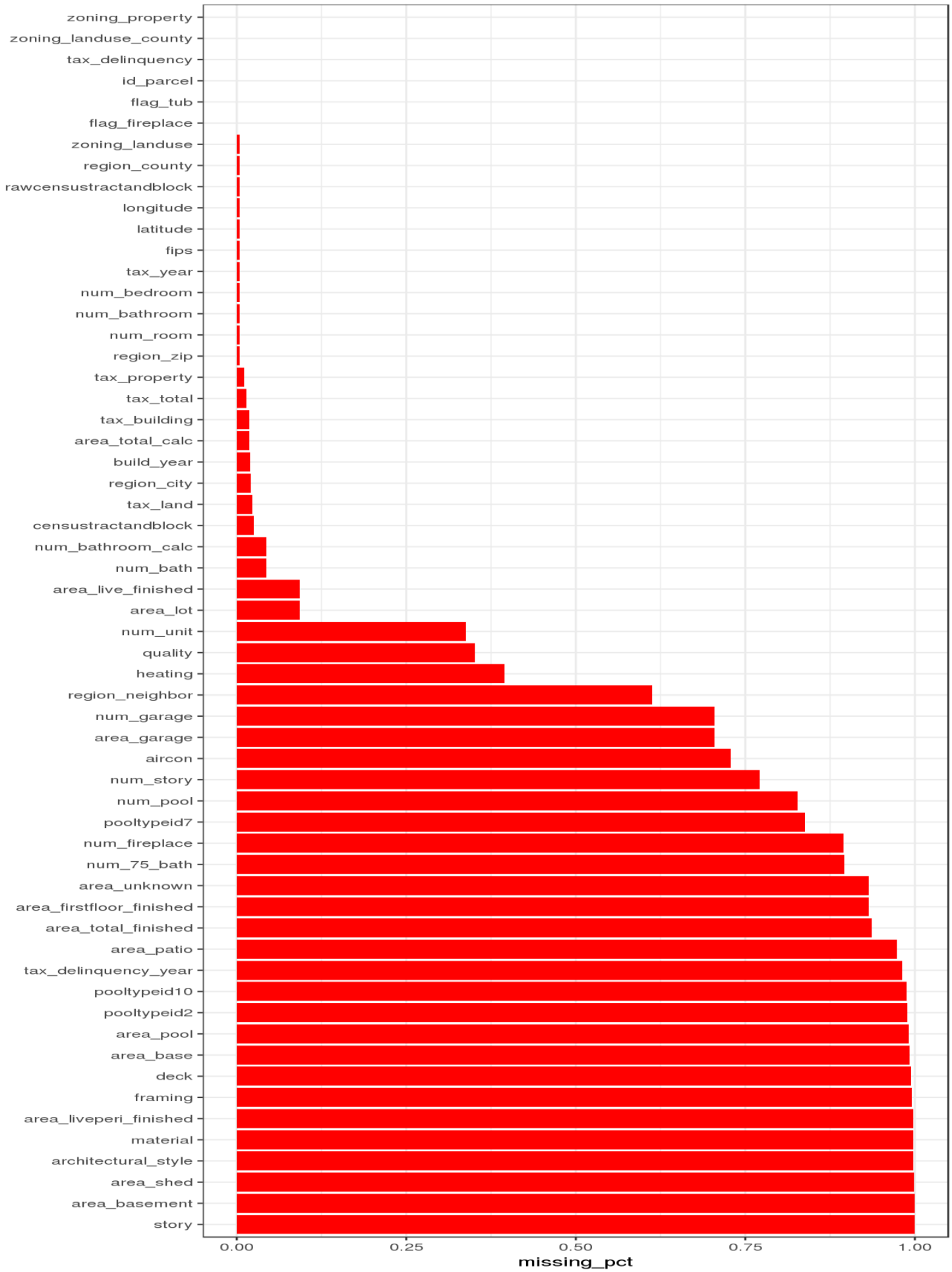
We will first do EDA & processing on the properties dataset:-

```
> dim(properties)
```

```
[1] 2985217      58
```

```
>t(head(properties,3))
```

parcelid	"10754147"	"10759547"	"10843547"
airconditioningtypeid	NA	NA	NA
architecturalstyletypeid	NA	NA	NA
basementsqft	NA	NA	NA
bathroomcnt	"0"	"0"	"0"
bedroomcnt	"0"	"0"	"0"
buildingclasstypeid	NA	NA	NA
buildingqualitytypeid	NA	NA	NA
calculatedbathnbr	NA	NA	NA
decktypeid	NA	NA	NA
finishedfloor1squarefeet	NA	NA	NA
calculatedfinishedsquarefeet	NA	NA	"73026"
finishedsquarefeet12	NA	NA	NA
finishedsquarefeet13	NA	NA	NA
finishedsquarefeet15	NA	NA	"73026"
finishedsquarefeet50	NA	NA	NA
finishedsquarefeet6	NA	NA	NA
fips	"6037"	"6037"	"6037"
fireplacecnt	NA	NA	NA
fullbathcnt	NA	NA	NA
garagecarcnt	NA	NA	NA
garagetotalsqft	NA	NA	NA
hashottuborspa	NA	NA	NA
heatingorsystemtypeid	NA	NA	NA
latitude	"34144442"	"34140430"	"33989359"
longitude	"-118654084"	"-118625364"	"-118394633"
lotsizesquarefeet	"85768"	" 4083"	"63085"
poolcnt	NA	NA	NA
poolsizesum	NA	NA	NA
pooltypeid10	NA	NA	NA
pooltypeid2	NA	NA	NA
pooltypeid7	NA	NA	NA
propertycountylandusecode	"010D"	"0109"	"1200"
propertylandusetypeid	"269"	"261"	" 47"
propertyzoningdesc	NA	"LCA11*"	"LAC2"
rawcensustractandblock	"60378002"	"60378001"	"60377030"
regionidcity	"37688"	"37688"	"51617"
regionidcounty	"3101"	"3101"	"3101"
regionidneighborhood	NA	NA	NA
regionidzip	"96337"	"96337"	"96095"
roomcnt	"0"	"0"	"0"
storytypeid	NA	NA	NA
threequarterbathnbr	NA	NA	NA
typeconstructiontypeid	NA	NA	NA
unitcnt	NA	NA	" 2"
yardbuildingsqft17	NA	NA	NA
yardbuildingsqft26	NA	NA	NA
yearbuilt	NA	NA	NA
numberofstories	NA	NA	NA
fireplaceflag	NA	NA	NA
structuretaxvaluedollarcnt	NA	NA	"650756"
taxvaluedollarcnt	" 9"	" 27516"	"1413387"
assessmentyear	"2015"	"2015"	"2015"
landtaxvaluedollarcnt	" 9"	" 27516"	"762631"
taxamount	NA	NA	"20800.37"
taxdelinquencyflag	NA	NA	NA
taxdelinquencyyear	NA	NA	NA
censustractandblock	NA	NA	NA


zoning_property
zoning_landuse_county
tax_delinquency
id_parcel
flag_tub
flag_fireplace
zoning_landuse
region_county
rawcensustractandblock
longitude
latitude
fips
tax_year
num_bedroom
num_bathroom
num_room
region_zip
tax_property
tax_total
tax_building
area_total_calc
build_year
region_city
tax_land
censustractandblock
num_bathroom_calc
num_bath
area_live_finished
area_lot
num_unit
quality
heating
region_neighbor
num_garage
area_garage
aircon
num_story
num_pool
pooltypeid7
num_fireplace
num_75_bath
area_unknown
area_firstfloor_finished
area_total_finished
area_patio
tax_delinquency_year
pooltypeid10
pooltypeid2
area_pool
area_base
deck
framing
area_liveperi_finished
material
architectural_style
area_shed
area_basement
story
0.00
0.25
0.50
0.75
1.00
missing_pct

The structure of the data is as shown above. The data is quite big with many variable & it can be noticed that most of them are NA.

Therefore the very first step of our analysis will be to calculate the missing value percentage present in every feature of the dataset. The graph shows the missing value percentage of features present in the dataset. As it is clear from the graph that most of the feature present have most of the values missing.

Therefore at the initial level only we shift our focus on the features which are not missing & thus we will be removing the features whose **missing percentage is greater than 60%.**
The missing value for the remaining feature were not carried out due to the large size of the data (approx.~ 30L) as it will lead to false imputation.

Now the next step will be to perform univariate analysis on the remaining feature of properties dataset.

UNIVARIATE ANALYSIS ON PROPERTY DATASET

After removing the features, 29 features are there which remained in the dataset.
Now each feature is analyzed individually to check the levels & understand that feature on the basis of information given in **zillow_data_dictionary.xlsx.**
The aim for this is to make sure that the class of the variable is not changed while loading the data & to change it to either categorical or continuous based on the **domain understanding.**

In this process, **Variable Consolidation** also take place by making sure to merging or removing the values/category which are acting as noise in the data.

VARIABLE CREATION ON PROPERTY DATASET

In the process , Certain new variables are created in addition to the feature present in the properties dataset. The feature are created by combination of two or more features & the knowledge of domain.

The Created Variables/Features are:-

1. **Bathcateg** – The category i.e. (S,M,L,XL) in which the property is divided on the basis of number of bathroom present in the dataset

2. **Bedcateg** - The category i.e. (S,M,L,XL) in which the property is divided on the basis of number of bedroom present in the dataset

3. **New_LivingAreaProp** – Proportion of living area present in the property.

   = calculatedfinishedsquarefeet / lotsizesquarefeet

4. **New_ValueProp** - Ratio of the built structure value to land area

   = structuretaxvaluedollarcnt / landtaxvaluedollarcnt

5. **New_ValueRatio** - Ratio of tax of property over parcel

   = taxvaluedollarcnt/ taxamount

6. **New_location** - addition of latitude & longitude coordinates of the data.
7. **New_ExtraSpace** - Amount of Reference Space

   = lotsizesquarefeet - calculatedfinishedsquarefeet
8. **New_TaxScore** - TotalTaxScore asssociating to the property

   = taxvaluedollarcnt * taxamount
9. **New_propertyage** - equal to the age of the property

   = 2017 - yearbuilt

Some of the features were also highly **co-related** to one another & thus one of them were dropped in that case on the basis of missing value percentage, the defination present in dictionary file & Domain Understanding.

The **corelated variable** were :-

- calculatedbathnbr & bathroomcnt
- finishedsquarefeet12 & calculatedfinishedsquarefeet
- bathroomcnt & fullbathcnt
- rawcensustractandblock & censustractandblock

TRANSACTION_DATASET:-

Now EDA is carried out on Transaction dataset (i.e. train_2016.csv) :-

```
> dim(transaction)
```

```
[1] 90275     3
```

```
> head(transaction)
```

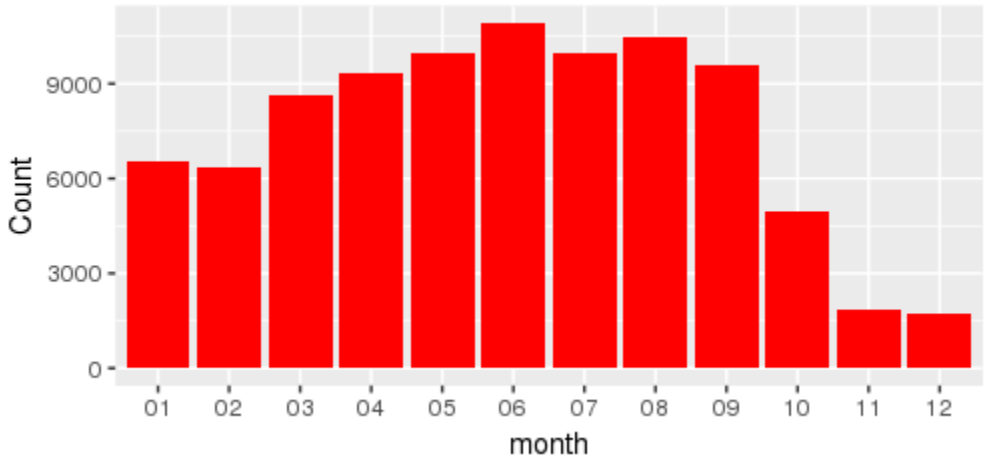
```
  parcelid logerror transactiondate
1 11016594   0.0276      2016-01-01
2 14366692  -0.1684      2016-01-01
3 12098116  -0.0040      2016-01-01
4 12643413   0.0218      2016-01-02
5 14432541  -0.0050      2016-01-02
6 11509835  -0.2705      2016-01-02
```

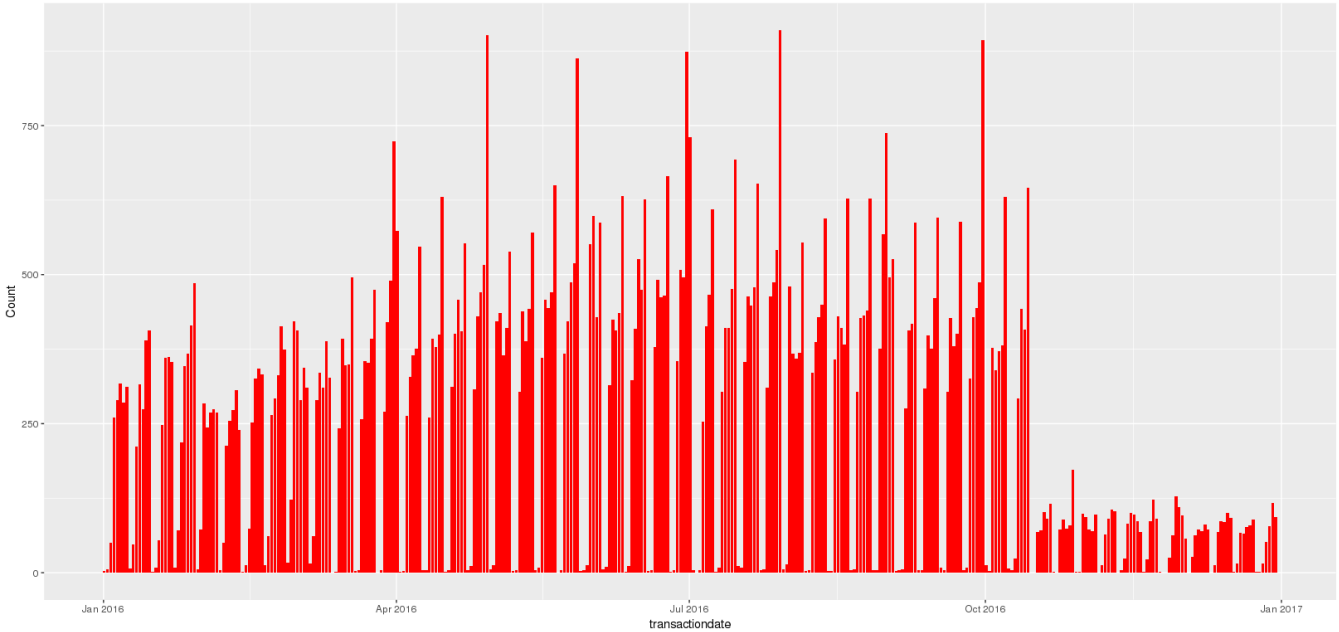
```
> summary(transaction)
```

```
   parcelid            logerror          transactiondate
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
```

There were **no missing value** present in the transaction dataset.

EXTRACTING MONTH FROM TRANSACTION DATE :- The next step was to understand that the month should be extracted from the transaction date as it was more insightful & the transaction carried out per month can thus be evaluated.





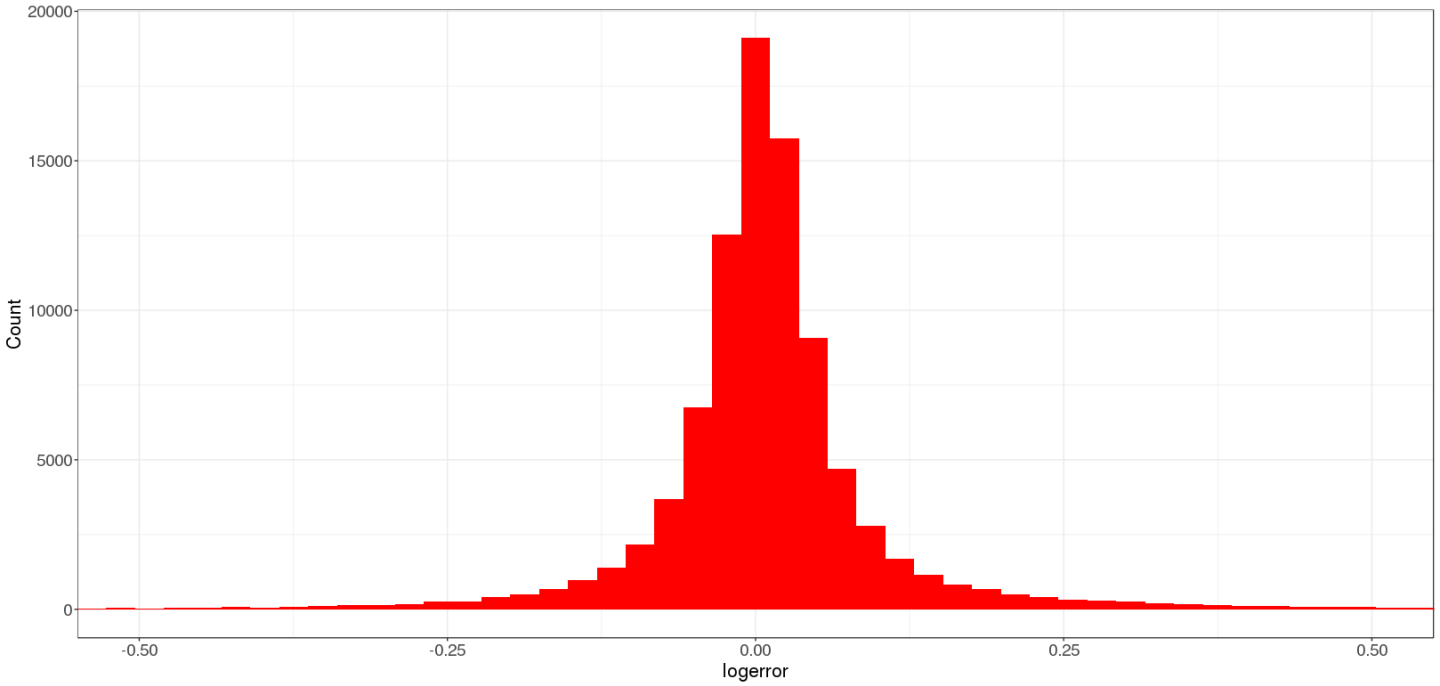
1st Graph :- **Transaction vs Month**

2nd Graph :- **Distribution of transaction dates**

LOG ERROR

Next up is the log error which is our target variable in this case study. Here the log error as predicted by the Zestimate for 90,150 properties are given (some parcel id are listed twice/ thrice indicating repeated resale of property)

Let us first visualize the distribution of log error.

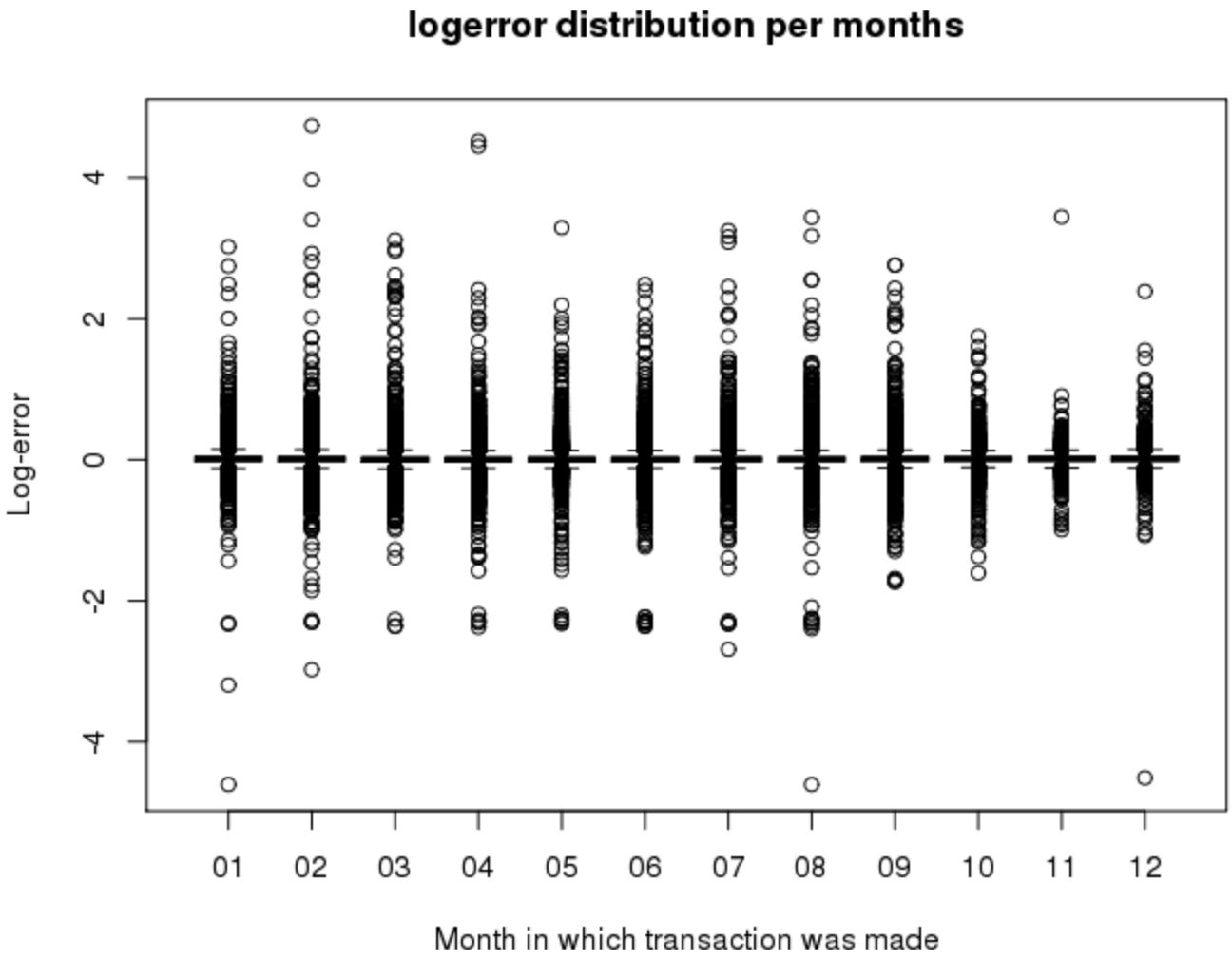


The Distribution of log error seems to be a nice normal distribution. There are some outliers present in the data because of which the data is spread widely over X-axis.

To understand the outlier distribution more clearly let us see the distribution of log error versus month.

Note :- In our whole study most of the outliers will not be treated even though a lot of outliers might be present. The reason for this is it will lead to wrong imputation or large missing data.Also in case of Housing problem the outliers may not be the result of mishandle of data.

Eg:- Sometimes homes are sold for a nominal amount to a close relative - say $1, leading to an overestimated Z. Other times an old home may be torn down, replaced and sold as a new structure and this could lead to an underestimated Z. So far I do not see how such anomalies can be detected/estimated based on the provided data.
Apart from that sometimes the factors such as inflation, property inheritance etc. lead to such results.



The above is the graph of log error vs the month of transaction.

4. EDA & DATA EXPLORATION ON MAIN TRAIN DATA

Main train data is the data which will be obtained after merging(left join) the transaction & properties data with “parcel id” as the key as it is present in both the datas.

```
prop_train = merge(transaction,properties,how = left,by = "parcelid")
```

This will lead to creation of a dataset which will be having all the information associated to houses in the transaction datasets along with the predicted logerror.

```
> dim(prop_train)
```

```
[1] 90275    61
```

MISSING VALUE IMPUTATION

There are many missing values present in the dataset but they can be predicted due to the reasonable size of the data. Mean and Median methods are checked for imputation but they are proved to be futile. KNN Imputation is much more appropriate for the data although it takes some time.

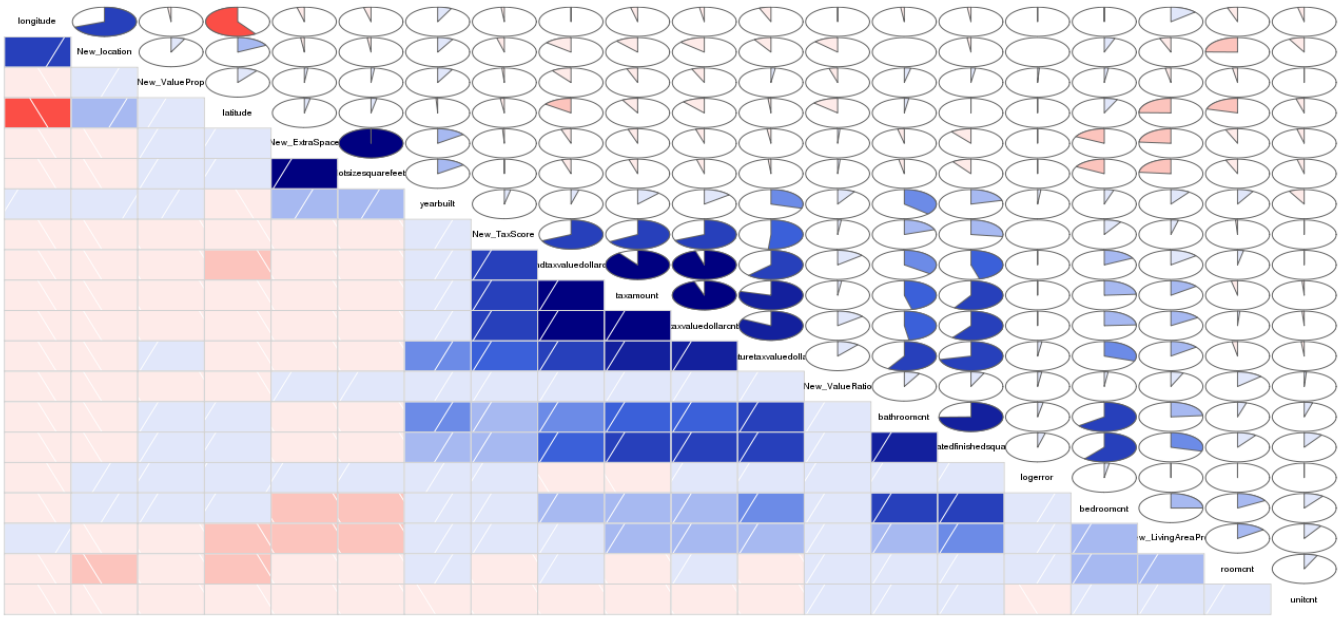
So the data is first divided into :-

- Factor_data :- data containing all the factor variables in the data
- Num :- Data Containing all the numerical variable of train

knnImputation function from the DMwR library is used for imputing the value.

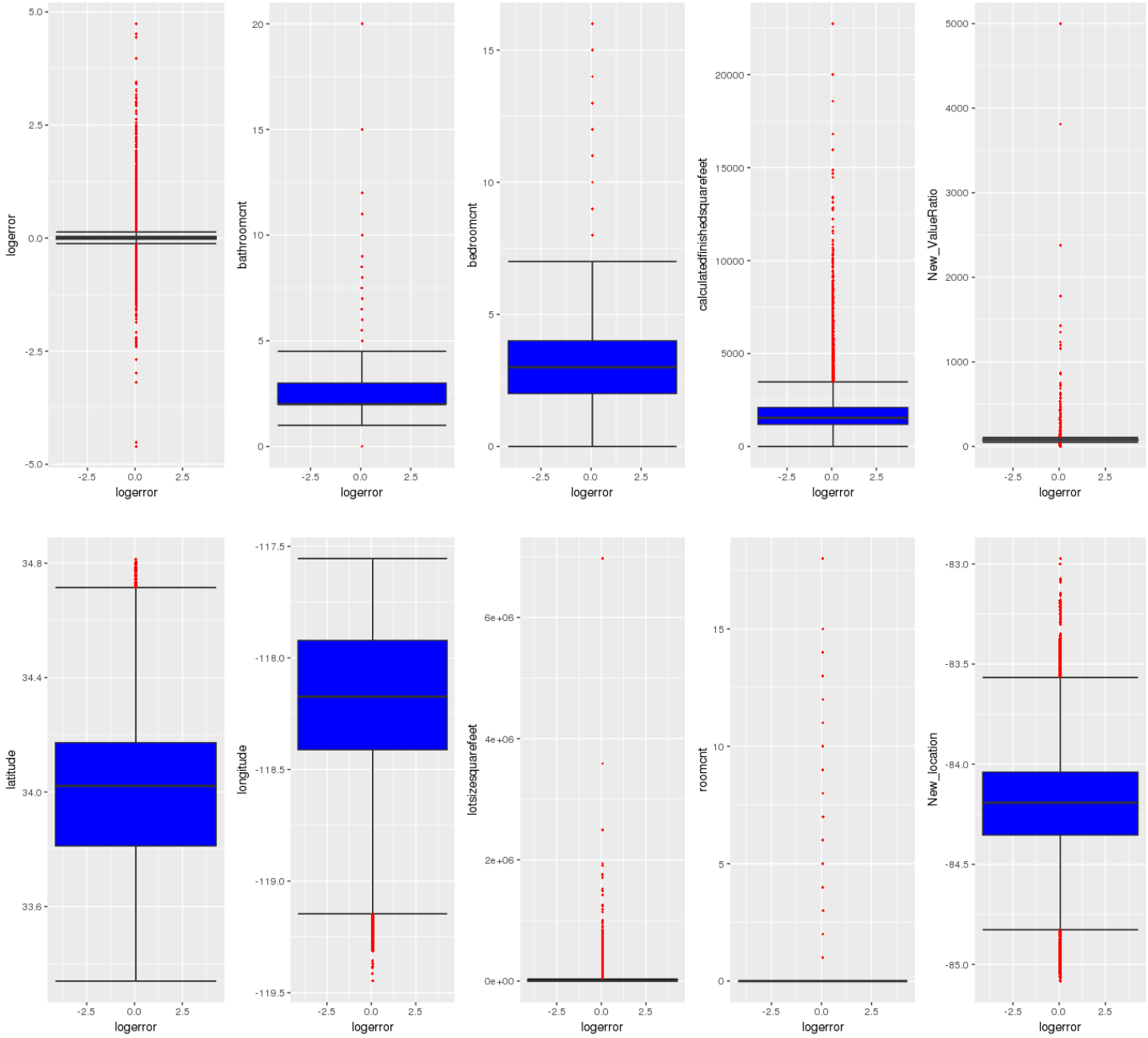
CORELATION ANALYSIS

```
> corrgram(num, order= T, upper.panel=panel.pie, text.panel=panel.txt, prop_train="Correlatio
n Plot")  # order = T causes PCA based reordering of the variables.
```



The corr-plot obtained is shown above. With the help of this plot highly correlated variable can be easily removed.

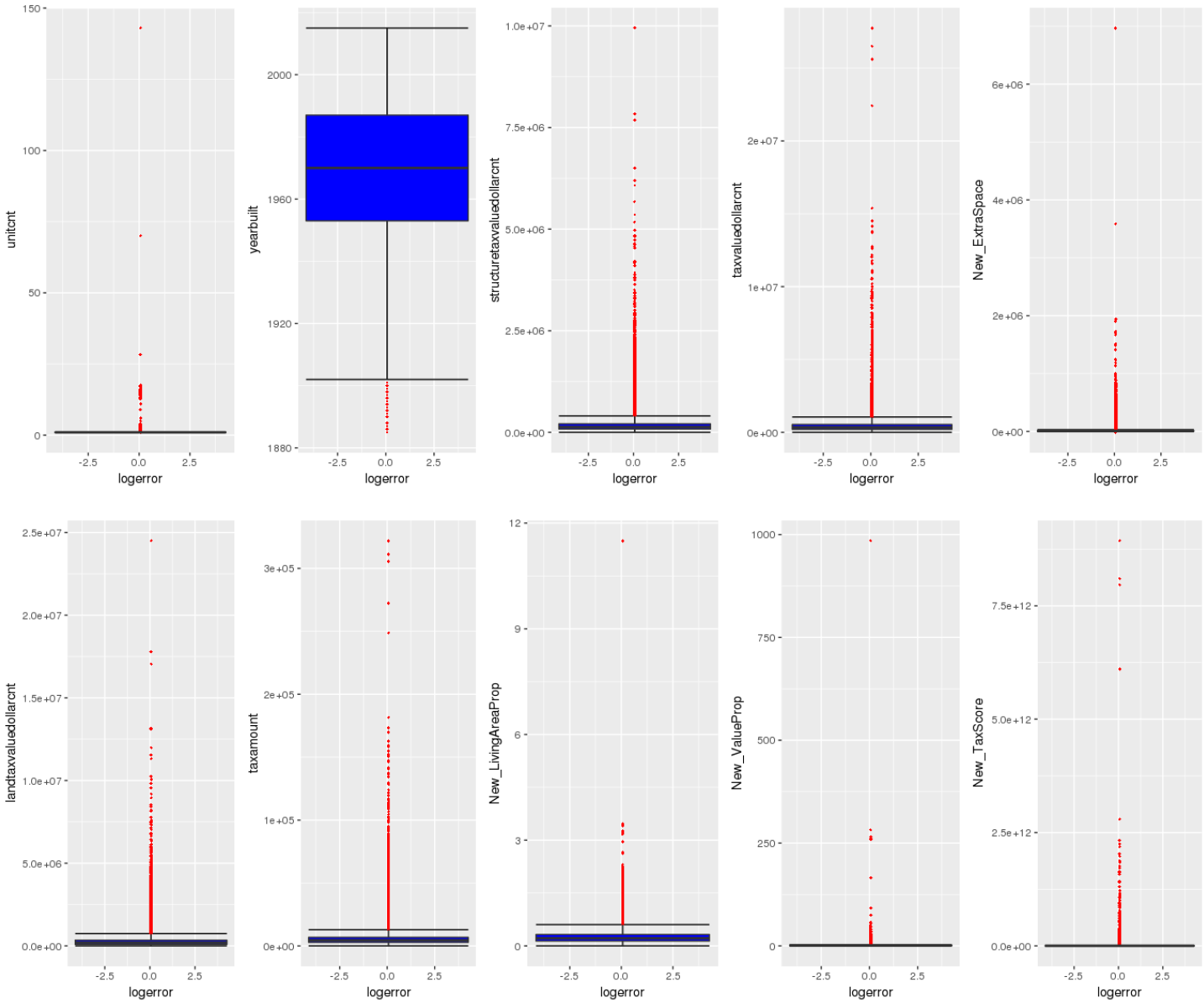
OUTLIER ANALYSIS



The box plot for the analysis of outlier are obtained & shown for all the variable vs the log error.

As can be seen there are many outliers present in the data but as covered in the note the outlier cannot be straight away treated as they may be genuine or will lead to loss of loads of data.

More data information is required so as to properly deal with the outliers.



CREATING TRAIN & TEST

We are now required to create the main test & train data so as that various model can be applied to these datasets.

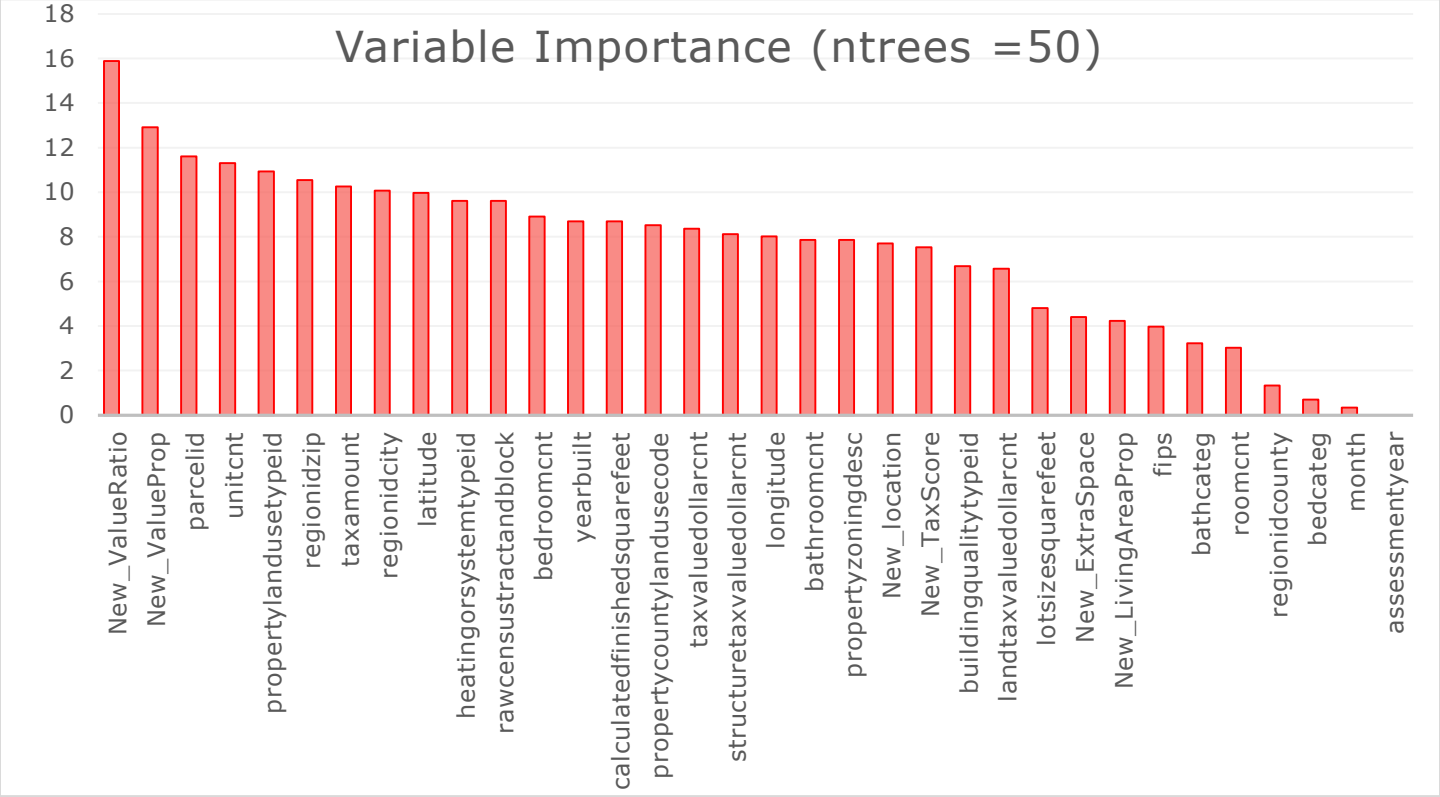
Train :- Left join of properties & transaction with imputed missing values & duplicates removed.

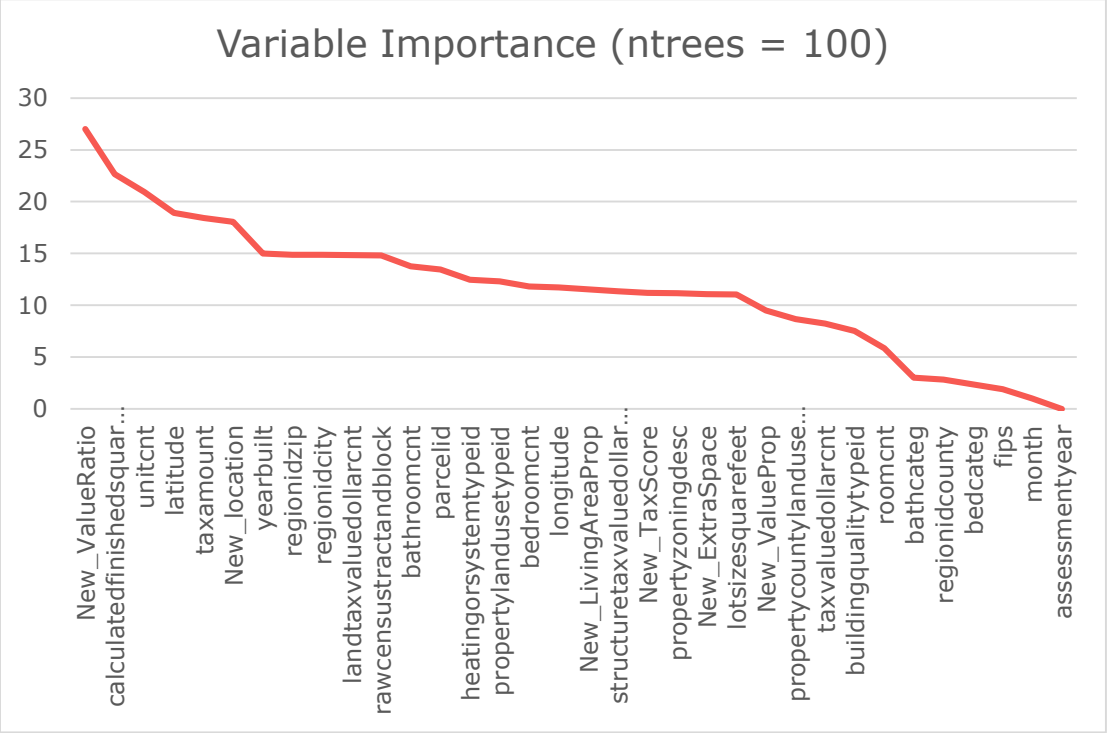
Test :- join of train & properties data with parcelid from train having all info. Logerror removed.

VARIABLE IMPORTANCE

Random Forest Algorithm was used for calculation of the Variable in the Train Dataset for the prediction of log error.

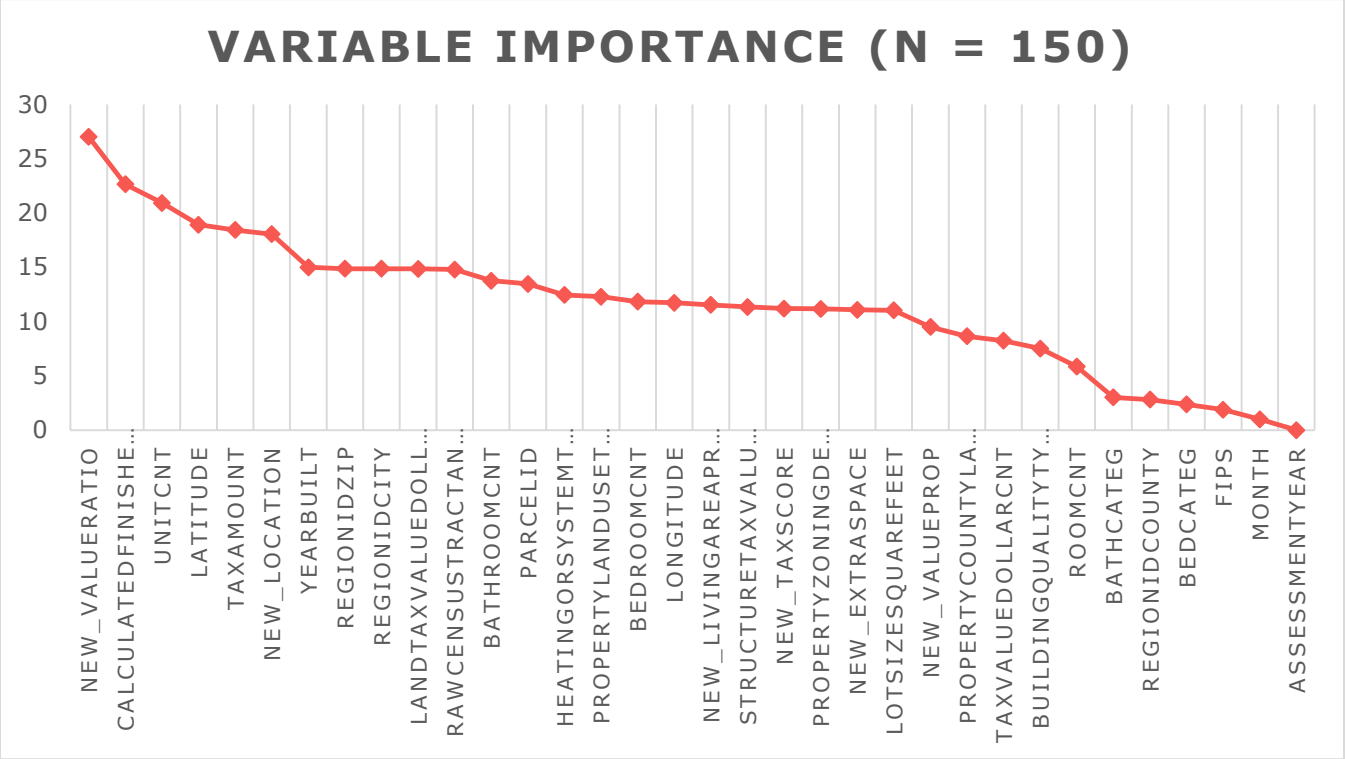
Variable importance was calculated for different no. of trees in the algorithm. (50 , 100 ,150)





It can be seen that the new features which were introduced in the dataset are greatly helping in explaining the dataset & are having good importance.

Further the importance of variables are changing with no. of trees. Although it can be easily seen that the top 15-20 features which are explaining the dataset are more or less same in every case.




```
>fit_regrex
```

```
  Call:
  randomForest(formula = logerror ~ ., data = main_train, ntree = 150,      importance = TRUE
,   n_jobs = -1)
               Type of random forest: regression
                     Number of trees: 150
  No. of variables tried at each split: 11

          Mean of squared residuals: 0.02584076
                    % Var explained: 0.41
```

MODEL BUILING

We are building our model using Statistical Methods and Machine learning algorithms for this analysis. Our case is different from a typical case of regression as in our case train data is a part of whole test data. Also information present in the test data is not given completely. Thus for this study we applied 5 machine learning algorithm as there are more chances of model failing due to above mentioned problem.

The Models for regression are :-

- Decision Trees
- Random Forest
- XGboost
- Linear regression
- Gradient boosting.[GBM]

As explained above all the models are applied for regression with their parameters tuned.

1. DECISION TREES

The very first model applied was decision trees for regression with method as “annova”.
The Data was applied first to train & then test data with logerror as the target.

Unfortunately Decision Tree Failed in predicting the log error for test data & for every parcel id same logerror was obtained.

Thus decision trees model was scraped off from our study.

2. RANDOM FOREST

Random forest model was initially used for the prediction of variable importance for different number of trees i.e. ntrees = 50,100,150. The same model was then used for the prediction of logerror on the test data.

RF failed in the regard that it was not able to predict majority of the logerror & labelled them as NA. Even on replacing the missing values in test data to 0 or other metric, same problem persist.

3. XGBOOST

XGboost model was trained on metric “mae” with the best possible value chosen for most of the parameters.
The model was trained & then applied to test for prediction of logerror of all houses in test.

The model gave a score of **0.1297686.**

4. SIMPLE LINEAR REGRESSION

Simple linear regression was applied in two ways. In the first case the model was trained on the dataset by including some of the most important features(ranging from 3,4 to 12-15)

In this case, the score ranged from about **0.0656812 to 0.0651196** which is a much greater improvement as compared to XG boost. The score varied greatly acc. To the features selected but was in the range 0.0656x.

The second way was when the model is trained on whole test data by taking into consideration only variables such as parcelid . fullbathcnt etc. i.e those having less missing value.
Then it was asked to predict on the same dataset. Interestingly the model performed quite well & gave a score of **0.0648837.**

5. GRADIENT BOOSTING (GBM)

The last model applied in our study was GBM due its power of prediction & superiority by which it can predict over other models.

All the train dataset was fed to the model & then the logerror was predicted.

The no. of trees selected was 200 while “Gaussian distribution” is selected. The other parameters were also optimally selected.

As expected GBM outperformed every single model & a score of **0.0648492** was obtained.

NOTE:- in each & every model the n_jobs were taken equal to -1 wherever possible so as to speed up the process & use maximum no. of cores available as the data is huge.

CONCLUSION

Zillow home prediction is the case where we were provided with a full list of real estate properties in three counties (Los Angeles, Orange and Ventura, California) data in 2016.In this type of housing problem a large number of factors based on location, area, geographical, social etc. have an effect on the price of the property. We were asked to predict the zestimate i.e.

$$\text{logerror}=\log(\text{Zestimate})-\log(\text{SalePrice})$$

We used a wide number of features given to us while we also applied domain knowledge to extract various other feature which can influence the estimated error. Also we showed through various method the importance of variables (both given & calculated).

Further we applied a large number of statistical & machine learning techniques so as to predict the mean error. Using the above we were able to predict with an absolute mean error of about **0.0648492.** These analysis will give consumers as much information as possible about homes and the housing market, marking the first time consumers had access to this type of home value information at no cost.