Capstone Project Report

By: Dice Analytics

Part 1: Definition

Project Overview:

The project which I have selected as capstone project of XXXX is the Zestimate Error prediction by the Kaggle. In the subject project we will be building a machine learning approach which will predict the log error for any of the property in the dataset. Ultimately it ll help Zillow to understand the reasons behind the error margin in Zillows prices prediction system.

For the prediction of the prices the Zillow have provided different information about houses like Number of rooms in the house, when the property was built, recent tax assessment, number of bathrooms in the property, furnished area, finished living area etc. These properties are well explained, quite general and explained in the above table.

Problem statement:

The goal is to create a prediction flow using machine learning algorithms to predict the logerror in the Zillow data set. During the problem solving following list of algorithm will be used to build up an approach which will come up with a minimum error against logerror value prediction. Further following steps will be followed to solve the problem.

1. Explanatory Data Analysis
2. Variate Analysis
3. Clustering using KMean/KModes
4. Prediction using Multiple Linear Regression

Metrics:

Since Zillow have provided the data to the public for understanding the reasons behind the log error & I have selected to the regression approach to minimize the logerror. So I will use R Squared Matric to track impact of different Machine Learning Approaches that how R Square will increase over the time & after applying different data mining approaches.

The purpose behind selecting the R Square as the metric is that I have selected an approach/framework, to address the problem where regressions are extensively being used. For regressions the R Square becomes a good choice to check the generalization of the model. Secondly, Zillow is looking for a way out using machine learning to minimize the log error rate in its prediction. So the R Sqaure will give us a fair idea that how good we can generalize the regression line. Once R Square enhanced then we can easily say that any other algorithm which will be used for the prediction of log error will perform well.

Generally, R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

The definition of R-squared is fairly straight-forward; it is the percentage of the response variable variation that is explained by a linear model. Or:

R-squared = Explained variation / Total variation

R-squared is always between 0 and 100%:

- 0% indicates that the model explains none of the variability of the response data around its mean.
- 100% indicates that the model explains all the variability of the response data around its mean.

In general, the higher the R-squared, the better the model fits your data means maximum variation within the data can be explained by the regression equation.

Part 2: Analysis

Data Exploration

On Kaggle website for the subject following type of files are provided

- Training file which contains only three features
  - **Parcel ID =>** ID Against Every Property
  - **Transaction Date =>** The date on which property was sold out
  - **Logerror =>** The Dependent or Target Variable, which explained by zilliow as log(predicted_price) – Log(actual_price)

- Properties file contains description about the property which was being traded. There are approximately 65 parameters in the properties file. Description about those features copied in the below table.

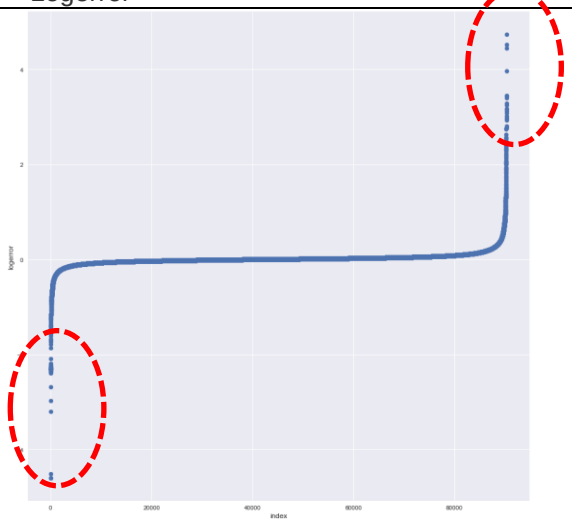
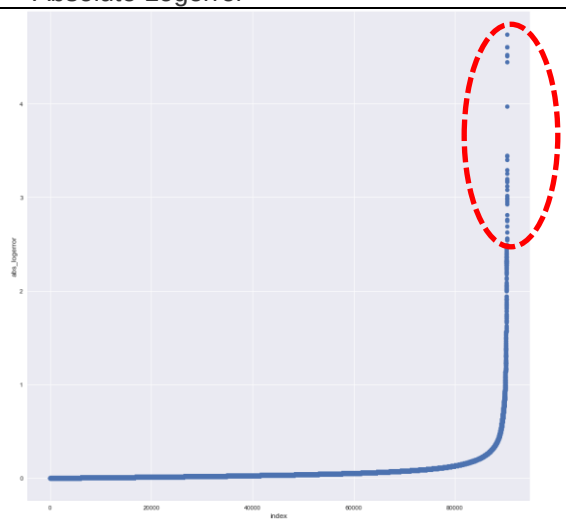
Feature	Description
'airconditioningtypeid'	Type of cooling system present in the home (if any)
'architecturalstyletypeid'	Architectural style of the home (i.e. ranch, colonial, split-level, etc…)
'basementsqft'	Finished living area below or partially below ground level
'bathroomcnt'	Number of bathrooms in home including fractional bathrooms
'bedroomcnt'	Number of bedrooms in home
'buildingqualitytypeid'	Overall assessment of condition of the building from best (lowest) to worst (highest)
'buildingclasstypeid'	The building framing type (steel frame, wood frame, concrete/brick)
'calculatedbathnbr'	Number of bathrooms in home including fractional bathroom
'decktypeid'	Type of deck (if any) present on parcel
'threequarterbathnbr'	Number of 3/4 bathrooms in house (shower + sink + toilet)
'finishedfloor1squarefeet'	Size of the finished living area on the first (entry) floor of the home
'calculatedfinishedsquarefeet'	Calculated total finished living area of the home
'finishedsquarefeet6'	Base unfinished and finished area
'finishedsquarefeet12'	Finished living area
'finishedsquarefeet13'	Perimeter living area
'finishedsquarefeet15'	Total area
'finishedsquarefeet50'	Size of the finished living area on the first (entry) floor of the home
'fips'	Federal Information Processing Standard code - see https://en.wikipedia.org/wiki/FIPS_county_code for more details
'fireplacecnt'	Number of fireplaces in a home (if any)
'fireplaceflag'	Is a fireplace present in this home
'fullbathcnt'	Number of full bathrooms (sink, shower + bathtub, and toilet) present in home
'garagecarcnt'	Total number of garages on the lot including an attached garage
'garagetotalsqft'	Total number of square feet of all garages on lot including an attached garage
'hashottuborspa'	Does the home have a hot tub or spa
'heatingorsystemtypeid'	Type of home heating system
'latitude'	Latitude of the middle of the parcel multiplied by 10e6
'longitude'	Longitude of the middle of the parcel multiplied by 10e6
'lotsizesquarefeet'	Area of the lot in square feet
'numberofstories'	Number of stories or levels the home has
'parcelid'	Unique identifier for parcels (lots)
'poolcnt'	Number of pools on the lot (if any)
'poolsizesum'	Total square footage of all pools on property
'pooltypeid10'	Spa or Hot Tub
'pooltypeid2'	Pool with Spa/Hot Tub
'pooltypeid7'	Pool without hot tub
'propertycountylandusecode'	County land use code i.e. it's zoning at the county level
'propertylandusetypeid'	Type of land use the property is zoned for
'propertyzoningdesc'	Description of the allowed land uses (zoning) for that property
'rawcensustractandblock'	Census tract and block ID combined - also contains blockgroup assignment by extension
'censustractandblock'	Census tract and block ID combined - also contains blockgroup assignment by extension
'regionidcounty'	County in which the property is located
'regionidcity'	City in which the property is located (if any)
'regionidzip'	Zip code in which the property is located
'regionidneighborhood'	Neighborhood in which the property is located
'roomcnt'	Total number of rooms in the principal residence
'storytypeid'	Type of floors in a multi-story house (i.e. basement and main level, split-level, attic, etc.). See tab for details.
'typeconstructiontypeid'	What type of construction material was used to construct the home
'unitcnt'	Number of units the structure is built into (i.e. 2 = duplex, 3 = triplex, etc...)
'yardbuildingsqft17'	Patio in yard
'yardbuildingsqft26'	Storage shed/building in yard
'yearbuilt'	The Year the principal residence was built
'taxvaluedollarcnt'	The total tax assessed value of the parcel
'structuretaxvaluedollarcnt'	The assessed value of the built structure on the parcel
'landtaxvaluedollarcnt'	The assessed value of the land area of the parcel
'taxamount'	The total property tax assessed for that assessment year
'assessmentyear'	The year of the property tax assessment
'taxdelinquencyflag'	Property taxes for this parcel are past due as of 2015
'taxdelinquencyyear'	Year for which the unpaid propert taxes were due

Exploratory Visualization

Explanatory Data Analysis is the key foundation layer of the any machine learning or data mining initiatives. So I have tried different aspects of Exploration Data Analysis to have a grip over the provided dataset, which will ultimately help me to come up with better machine learning approaches.

Range is an important property of a feature which gives information about the min & max values in the dataset. So I sorted the logerror in descending order and plotted given below graph to have a sneak peek

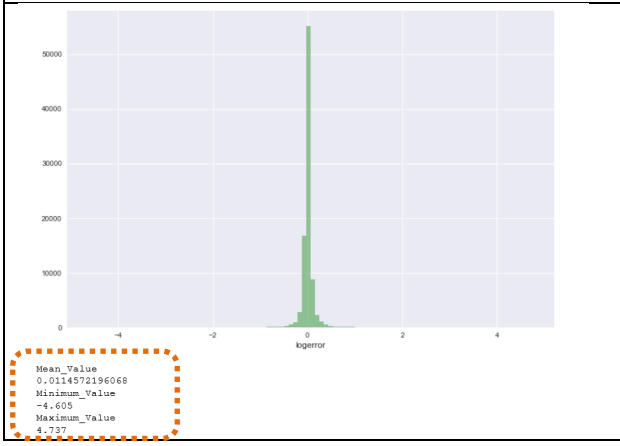
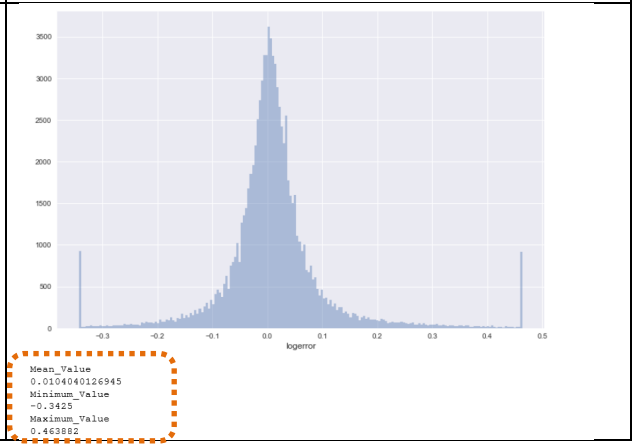
into logerror. After looking at the graphs it's quite clear that there are out-liers in the logerror on both negative logerror side and positive side error & the picture get more clear when we draw absolute logerror.

Logerror	Absolute Logerror
	

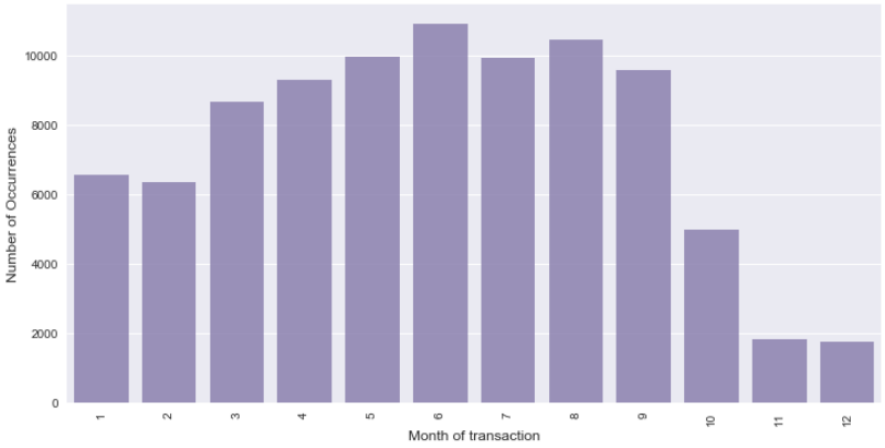
As a technique to remove **outliers**, we divided data into 100 equal parts and Lowest 1% of rows wrt. Logerror and top 1% with respect to logerror were removed from the dataset i.e. outliers were removed. This technique resembles with the outliers removal using standard deviation & we removed the rows in dataset which were +/- 3 Standard deviation away from mean. Please find below the snapshot of python code used for the outliers removal

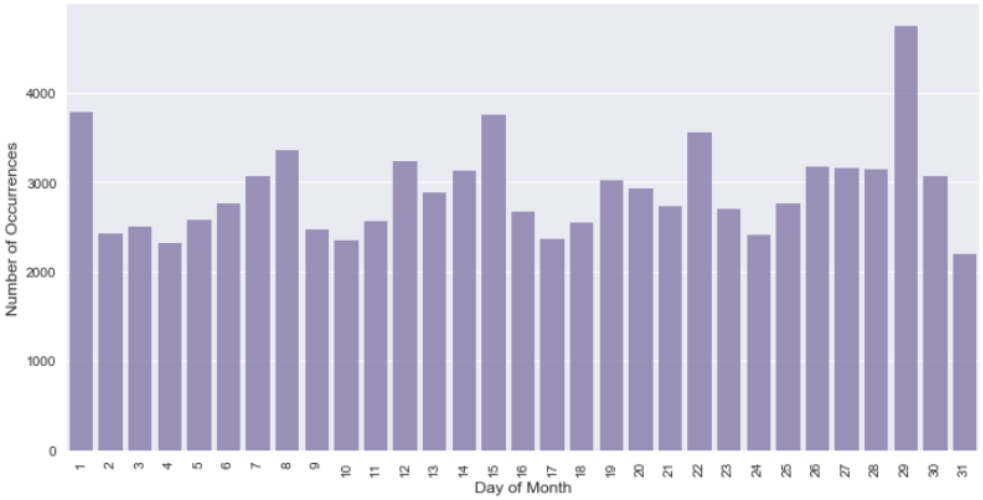
```
ulimit = np.percentile(training_data.logerror.values, 99)
llimit = np.percentile(training_data.logerror.values, 1)
training_data['logerror'].ix[training_data['logerror']>ulimit] = ulimit
training_data['logerror'].ix[training_data['logerror']<llimit] = llimit
```

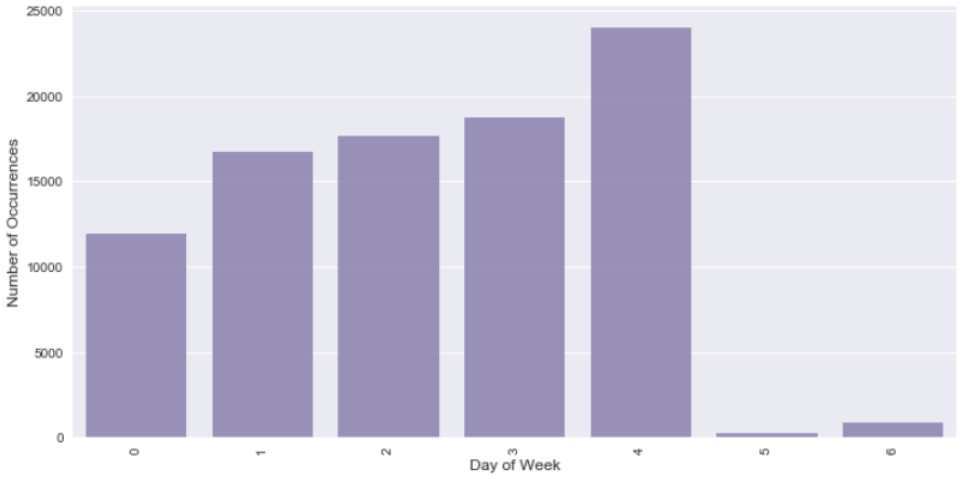
After the removal of the 2% of data from the given dataset following data distribution emerged.

Before outliers removal (Logerror)	**After outliers removal (Logerror)**
	

Now let’s explore the transaction date feature to have an understanding of the sales trend. Please find below few of the graphs to have an understanding of the sales trend w.r.t. to Month, Week of year, Day of week etc.





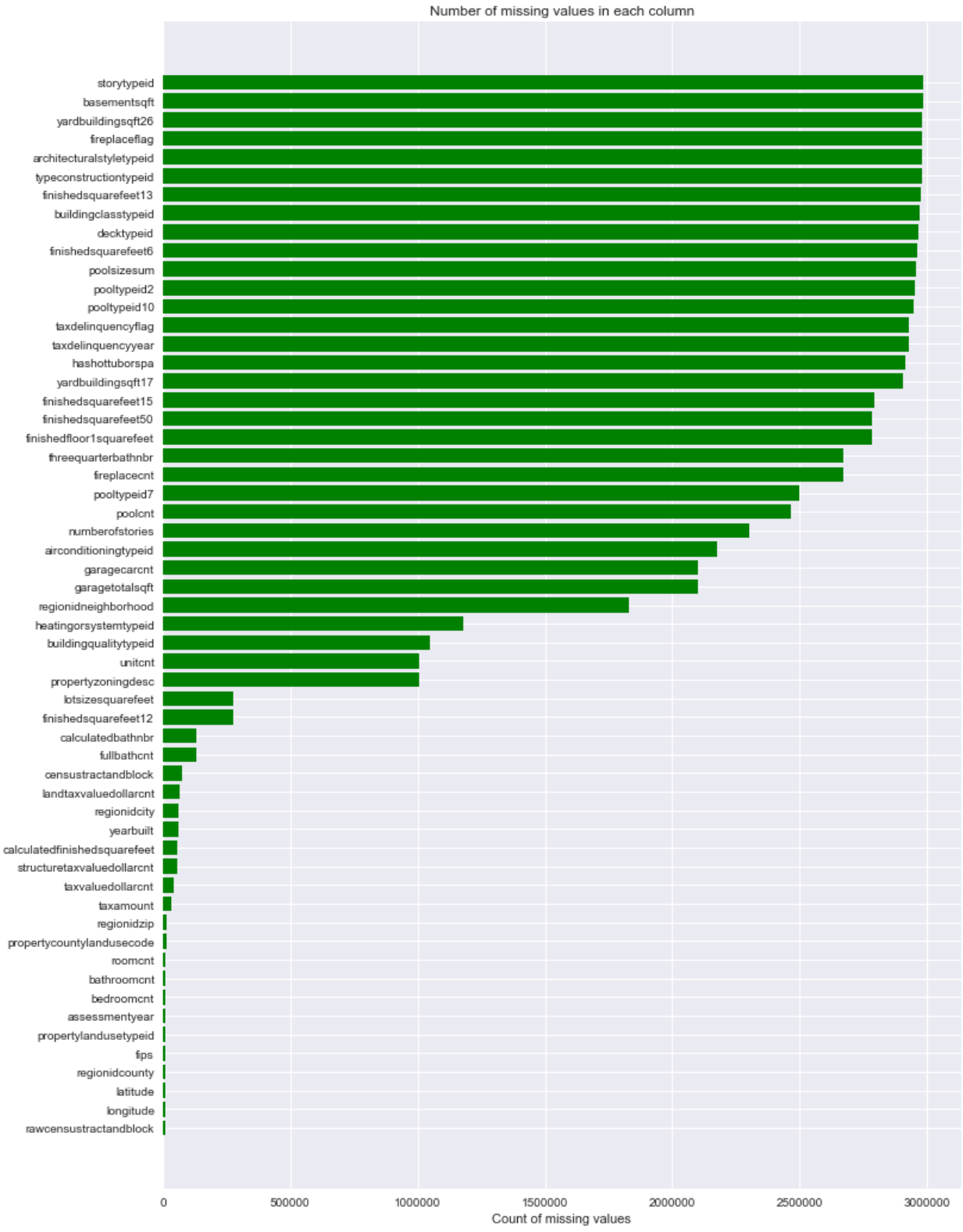


After looking at the above graphs following insights were observed.

1. The provided dataset seems to be containing most of the transactions from the first nine (09) months of the 2016.
2. Start of month and end of any month seems to be the day, where sale and purchase of the property is usually quite high
3. 5th & 6th day of every week i.e. Saturday & Sunday seems to be close business days as per given data that's why real-state trade seems quite low on the days.

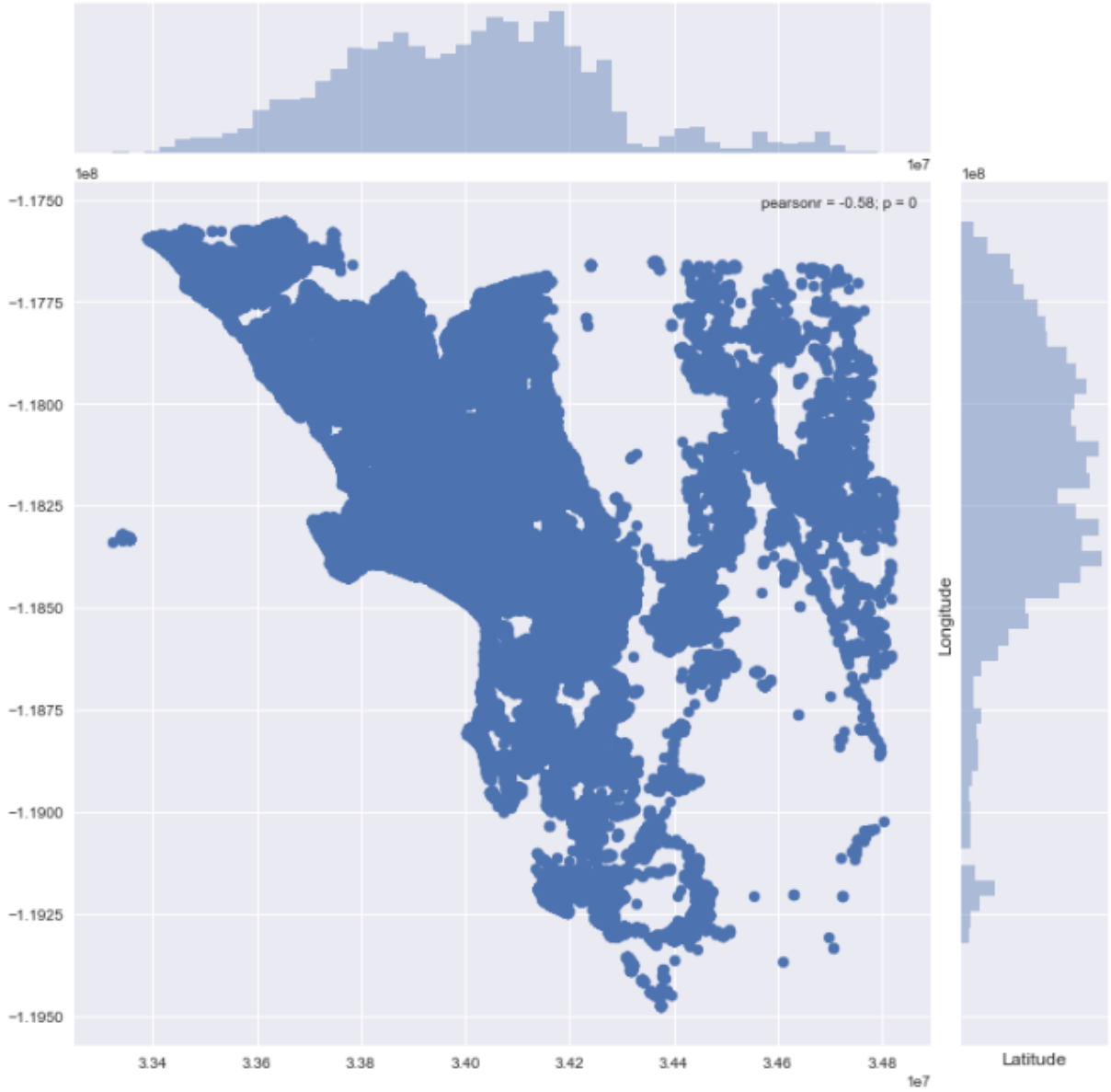
Missing values are an important aspect of the any given data set. Because ratio of the missing values let us know how deliberately we can rely on a feature. PFB table and graphs for portraying the missing values in the properties file provided by Zillow

	column_name	missing_count	missing_ratio
2	architecturalstyletypeid	2979156	0.997970
3	basementsqft	2983589	0.999455
6	buildingclasstypeid	2972588	0.995769
9	decktypeid	2968121	0.994273
10	finishedfloor1squarefeet	2782500	0.932093
13	finishedsquarefeet13	2977545	0.997430
14	finishedsquarefeet15	2794419	0.936086
15	finishedsquarefeet50	2782500	0.932093
16	finishedsquarefeet6	2963216	0.992630
18	fireplacecnt	2672580	0.895272
22	hashottuborspa	2916203	0.976881
27	poolcnt	2467683	0.826634
28	poolsizesum	2957257	0.990634
29	pooltypeid10	2948278	0.987626
30	pooltypeid2	2953142	0.989255
31	pooltypeid7	2499758	0.837379
41	storytypeid	2983593	0.999456
42	threequarterbathnbr	2673586	0.895609
43	typeconstructiontypeid	2978470	0.997740
45	yardbuildingsqft17	2904862	0.973082
46	yardbuildingsqft26	2982570	0.999113
49	fireplaceflag	2980054	0.998270
55	taxdelinquencyflag	2928755	0.981086
56	taxdelinquencyyear	2928753	0.981085


Number of missing values in each column
storytypeid
basementsqft
yardbuildingsqft26
fireplaceflag
architecturalstyletypeid
typeconstructiontypeid
finishedsquarefeet13
buildingclasstypeid
decktypeid
finishedsquarefeet6
poolsizesum
pooltypeid2
pooltypeid10
taxdelinquencyflag
taxdelinquencyyear
hashottuborspa
yardbuildingsqft17
finishedsquarefeet15
finishedsquarefeet50
finishedfloor1squarefeet
threequarterbathnbr
fireplacecnt
pooltypeid7
poolcnt
numberofstories
airconditioningtypeid
garagecarcnt
garagetotalsqft
regionidneighborhood
heatingorsystemtypeid
buildingqualitytypeid
unitcnt
propertyzoningdesc
lotsizesquarefeet
finishedsquarefeet12
calculatedbathnbr
fullbathcnt
censustractandblock
landtaxvaluedollarcnt
regionidcity
yearbuilt
calculatedfinishedsquarefeet
structuretaxvaluedollarcnt
taxvaluedollarcnt
taxamount
regionidzip
propertycountylandusecode
roomcnt
bathroomcnt
bedroomcnt
assessmentyear
propertylandusetypeid
fips
regionidcounty
latitude
longitude
rawcensustractandblock
0
500000
1000000
1500000
2000000
2500000
3000000
Count of missing values

After looking at the missing values it seems that there are alot of features where missing values percentage is quite high even reaches to 99%. And moving forward we should think about replacement of the missing values.

In the world of property & realstate, the location aspect acts as one of the major driving factor to determine the price of the real state property. Below graph represents the geo distribution of the provided data. It's quite clear from the below graphs that most of the properties belongs to a dense area. While there seems to be another cluster within the provided data



Now the next step is **to join both the properties and training file** to have a complete picture of the training dataset. Previous data explorations were performed on the properties files alone.

Now after joining with the training dataset, PFB the update distribution of the missing values

Before joining with the training dataset

	column_name	missing_count	missing_ratio
2	architecturalstyletypeid	2979156	0.997970
3	basementsqft	2983589	0.999455
6	buildingclasstypeid	2972588	0.995769
9	decktypeid	2968121	0.994273
10	finishedfloor1squarefeet	2782500	0.932093
13	finishedsquarefeet13	2977545	0.997430
14	finishedsquarefeet15	2794419	0.936086
15	finishedsquarefeet50	2782500	0.932093
16	finishedsquarefeet6	2963216	0.992630
18	fireplacecnt	2672580	0.895272
22	hashottuborspa	2916203	0.976881
27	poolcnt	2467683	0.826634
28	poolsizesum	2957257	0.990634
29	pooltypeid10	2948278	0.987626
30	pooltypeid2	2953142	0.989255
31	pooltypeid7	2499758	0.837379
41	storytypeid	2983593	0.999456
42	threequarterbathnbr	2673586	0.895609
43	typeconstructiontypeid	2978470	0.997740
45	yardbuildingsqft17	2904862	0.973082
46	yardbuildingsqft26	2982570	0.999113
49	fireplaceflag	2980054	0.998270
55	taxdelinquencyflag	2928755	0.981086
56	taxdelinquencyyear	2928753	0.981085

After joining with the training dataset

	column_name	missing_count	missing_ratio
5	architecturalstyletypeid	90014	0.997109
6	basementsqft	90232	0.999524
9	buildingclasstypeid	90259	0.999823
12	decktypeid	89617	0.992711
13	finishedfloor1squarefeet	83419	0.924054
16	finishedsquarefeet13	90242	0.999634
17	finishedsquarefeet15	86711	0.960521
18	finishedsquarefeet50	83419	0.924054
19	finishedsquarefeet6	89854	0.995336
21	fireplacecnt	80668	0.893581
25	hashottuborspa	87910	0.973802
30	poolcnt	72374	0.801706
31	poolsizesum	89306	0.989266
32	pooltypeid10	89114	0.987139
33	pooltypeid2	89071	0.986663
34	pooltypeid7	73578	0.815043
44	storytypeid	90232	0.999524
45	threequarterbathnbr	78266	0.866973
46	typeconstructiontypeid	89976	0.996688
48	yardbuildingsqft17	87629	0.970690
49	yardbuildingsqft26	90180	0.998948
52	fireplaceflag	90053	0.997541
58	taxdelinquencyflag	88492	0.980249
59	taxdelinquencyyear	88492	0.980249

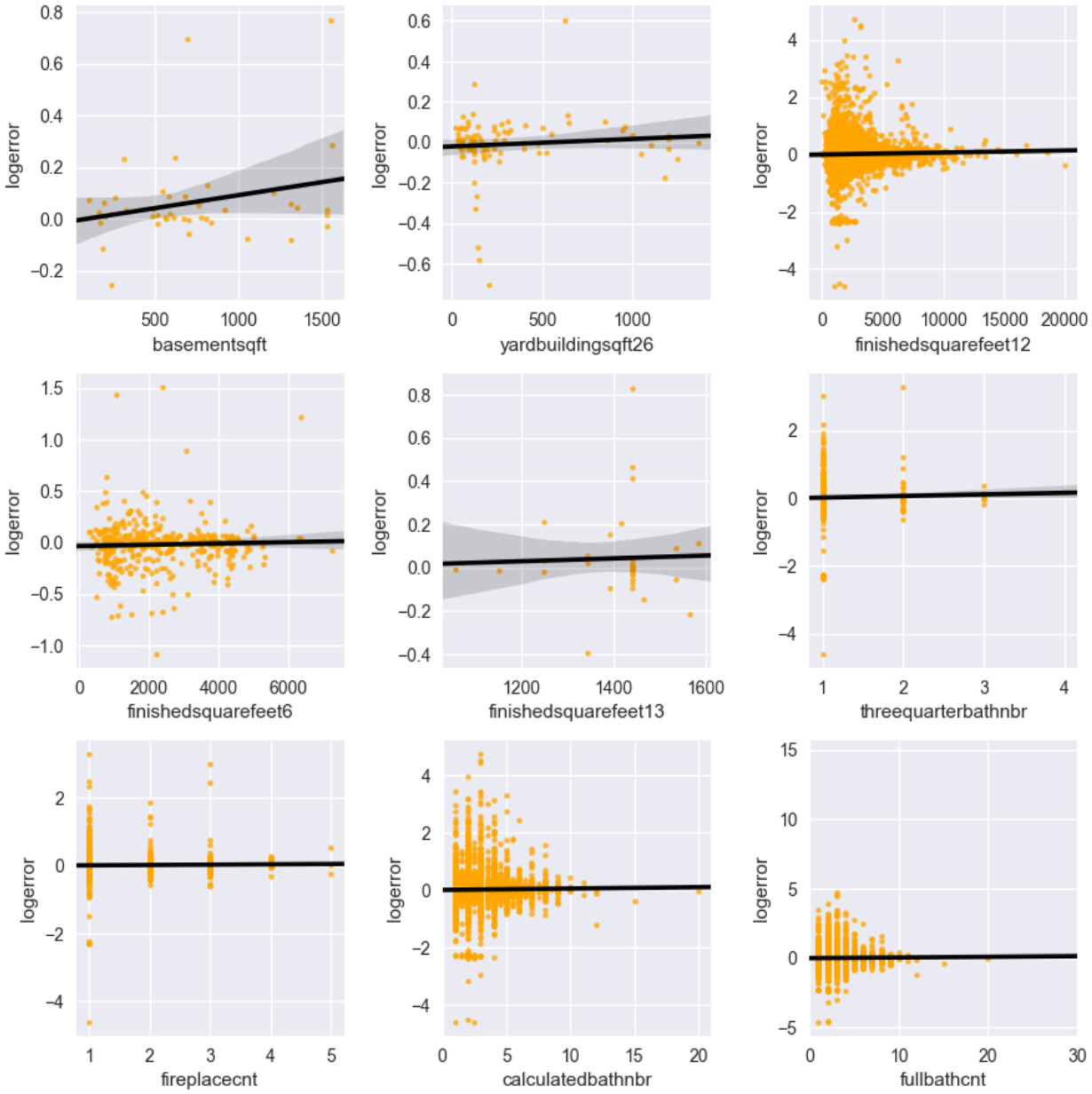
After reading through the available table it seems that situation of missing values even exists for training dataset. i.e. the same proportion of missing values remains

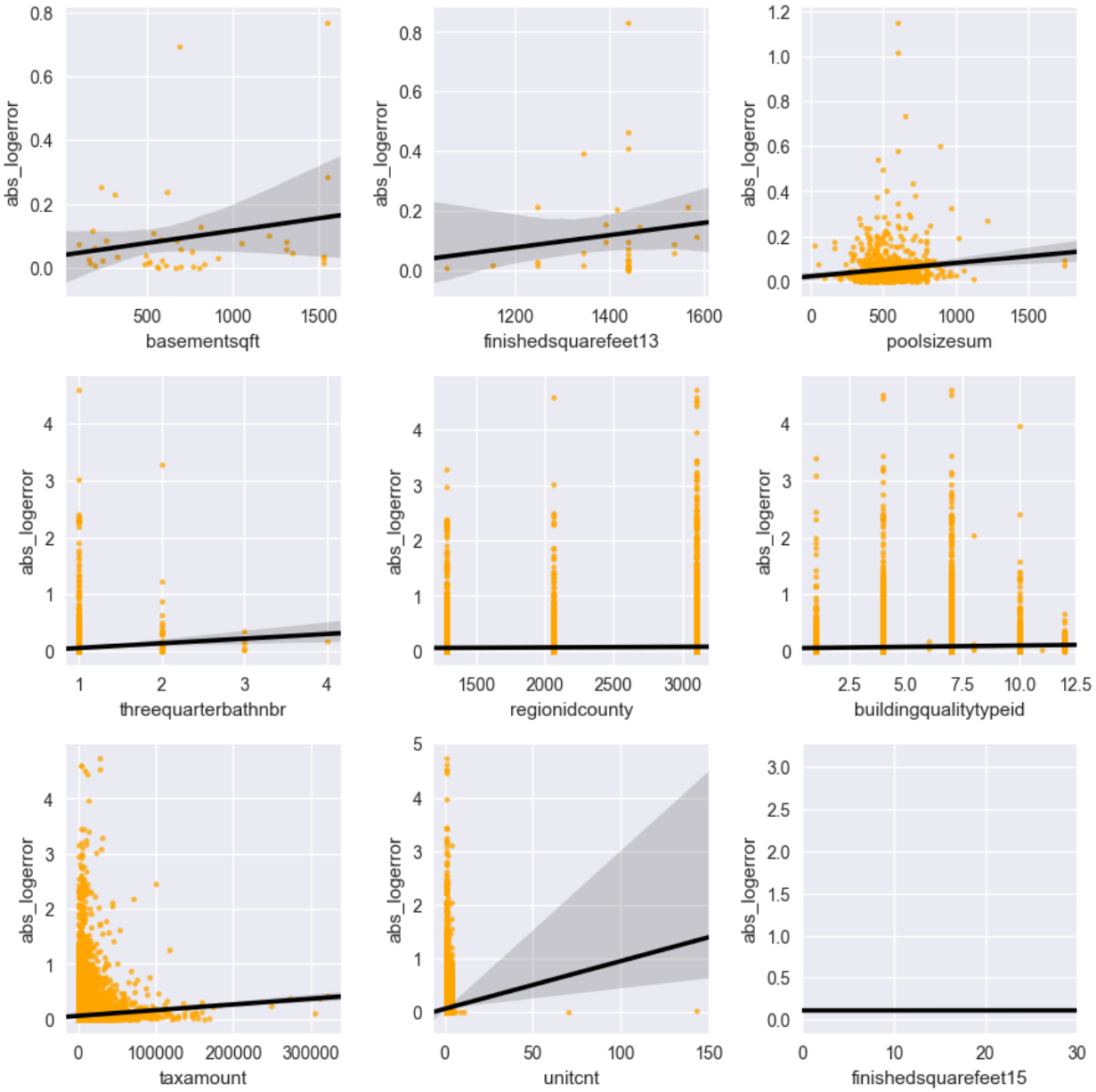
Now the next step is to build the **correlation matrix** w.r.t. to log error and absolute log error that which of the particular feature or independent feature have maximum information related to the log error or absolute logerror. After calculating the correlation matrix, there seems to be emerged an interesting fact that the correlations with the Absolute Logerror feature are relative higher as compared to logerror. In the correlation table with the logerror its quite clear that only one variable have correlation higher than 0.1 while in absolute log error table there are atleast 3 variables which have correlation higher than 0.1 in the correlation table.

Correlation with the Logerror	Correlation with Absolute Logerror

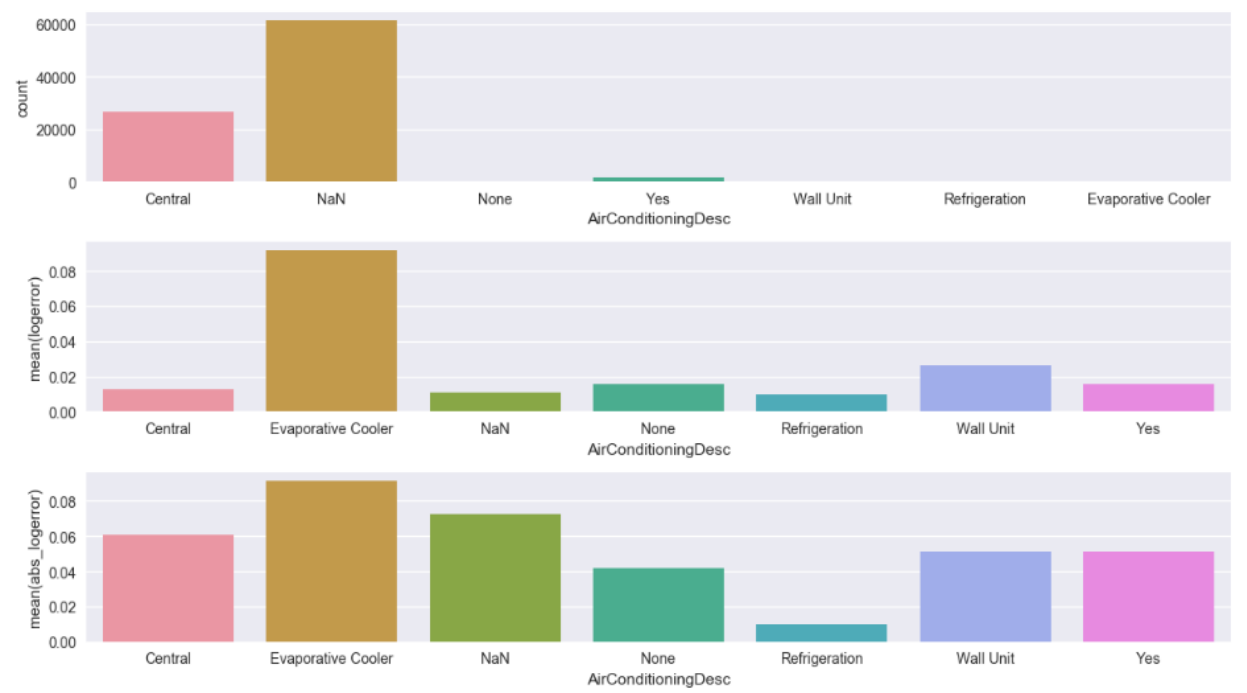
basementsqft	0.253395	basementsqft	0.212579
yardbuildingsqft26	0.086179	finishedsquarefeet13	0.129071
finishedsquarefeet12	0.041922	poolsizesum	0.109684
calculatedfinishedsquarefeet	0.038784	threequarterbathnbr	0.064325
finishedsquarefeet6	0.036847	regionidcounty	0.061964
finishedsquarefeet13	0.034715	buildingqualitytypeid	0.059824
threequarterbathnbr	0.034069	taxamount	0.048277
fireplacecnt	0.033235	unitcnt	0.046517
calculatedbathnbr	0.029448	finishedsquarefeet15	0.045420
fullbathcnt	0.028845	heatingorsystemtypeid	0.040220

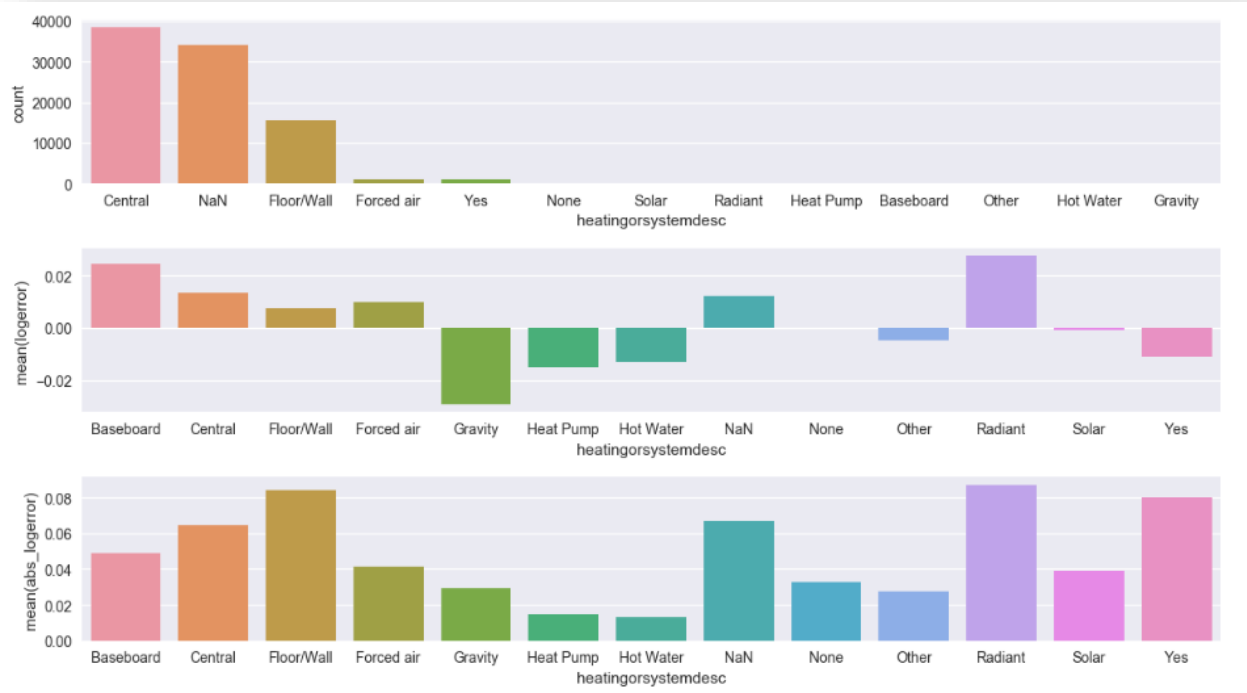
Same correlation/relationship of independent variables with the dependent variables can be cross checked through regression plots.

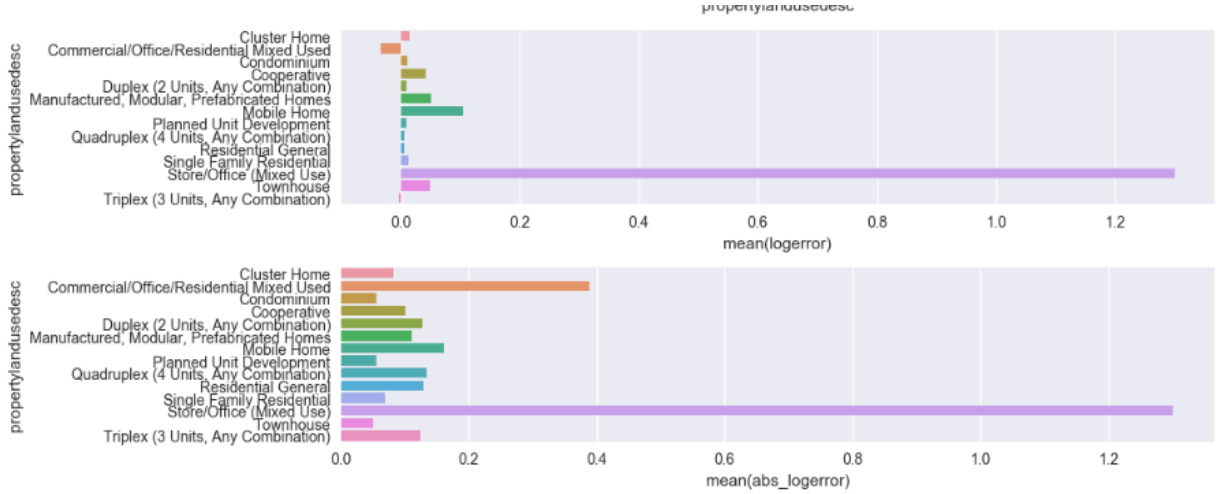


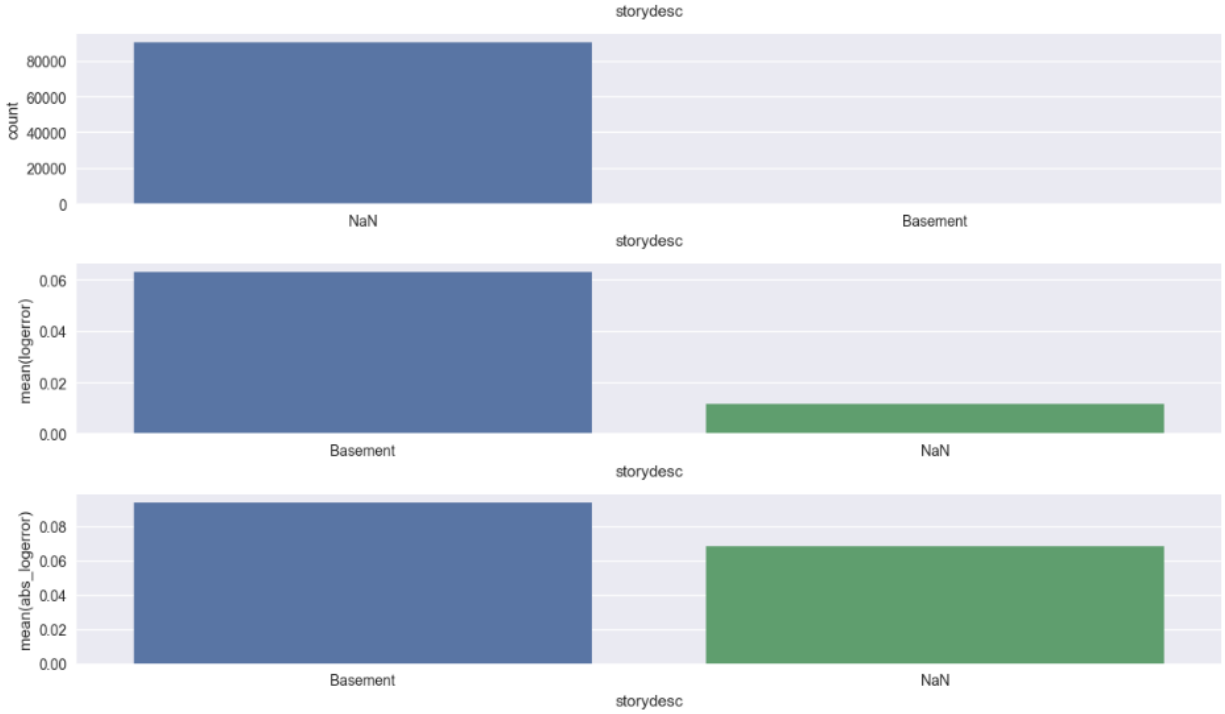


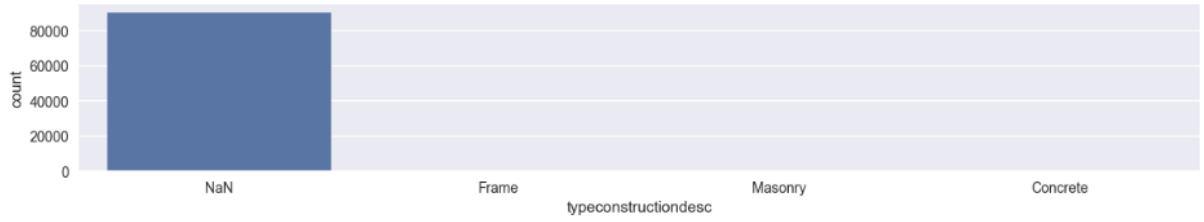
Now let's start inspecting the categorical data as a lot of variables which are part of the training set are categorical variables. We will try to build a relationship with the log error and absolute log error.

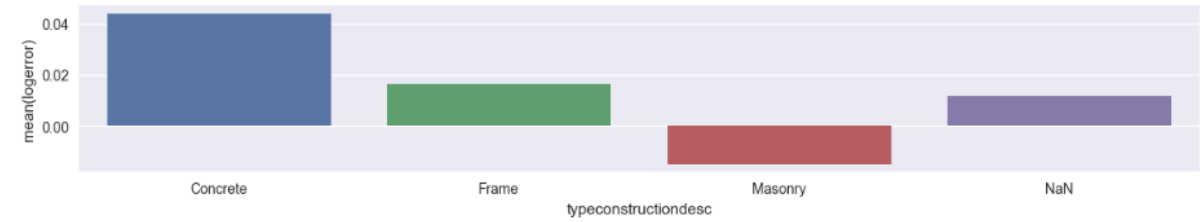

count
60000
40000
20000
0
Central
NaN
None
Yes
Wall Unit
Refrigeration
Evaporative Cooler
AirConditioningDesc
mean(logerror)
0.08
0.06
0.04
0.02
0.00
Central
Evaporative Cooler
NaN
None
Refrigeration
Wall Unit
Yes
AirConditioningDesc
mean(abs_logerror)
0.08
0.06
0.04
0.02
0.00
Central
Evaporative Cooler
NaN
None
Refrigeration
Wall Unit
Yes
AirConditioningDesc

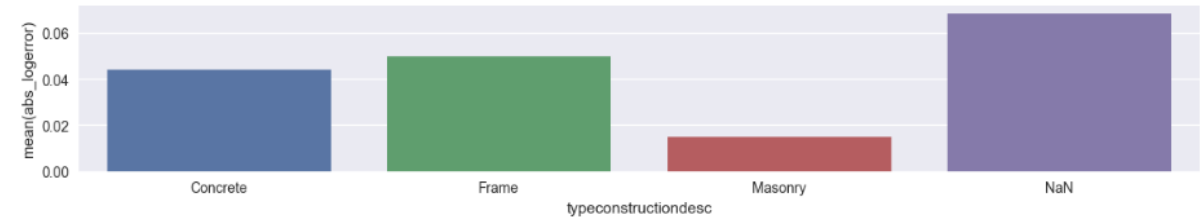

count
40000
30000
20000
10000
0
Central
NaN
Floor/Wall
Forced air
Yes
None
Solar
Radiant
Heat Pump
Baseboard
Other
Hot Water
Gravity
heatingorsystemdesc
mean(logerror)
0.02
0.00
−0.02
Baseboard
Central
Floor/Wall
Forced air
Gravity
Heat Pump
Hot Water
NaN
None
Other
Radiant
Solar
Yes
heatingorsystemdesc
mean(abs_logerror)
0.08
0.06
0.04
0.02
0.00
Baseboard
Central
Floor/Wall
Forced air
Gravity
Heat Pump
Hot Water
NaN
None
Other
Radiant
Solar
Yes
heatingorsystemdesc

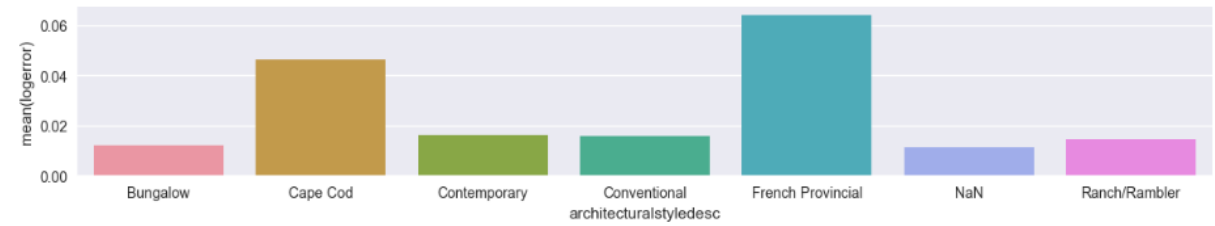

propertylandusedesc
Cluster Home
Commercial/Office/Residential Mixed Used
Condominium
Cooperative
Duplex (2 Units, Any Combination)
Manufactured, Modular, Prefabricated Homes
Mobile Home
Planned Unit Development
Quadruplex (4 Units, Any Combination)
Residential General
Single Family Residential
Store/Office (Mixed Use)
Townhouse
Triplex (3 Units, Any Combination)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
mean(logerror)
propertylandusedesc
Cluster Home
Commercial/Office/Residential Mixed Used
Condominium
Cooperative
Duplex (2 Units, Any Combination)
Manufactured, Modular, Prefabricated Homes
Mobile Home
Planned Unit Development
Quadruplex (4 Units, Any Combination)
Residential General
Single Family Residential
Store/Office (Mixed Use)
Townhouse
Triplex (3 Units, Any Combination)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
mean(abs_logerror)

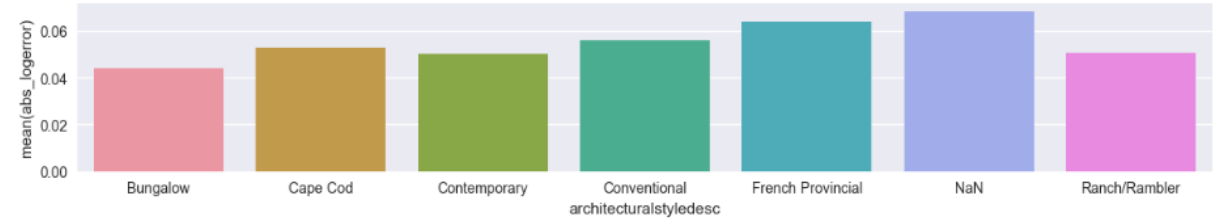

storydesc
count
80000
60000
40000
20000
0
NaN
Basement
storydesc
mean(logerror)
0.06
0.04
0.02
0.00
Basement
NaN
storydesc
mean(abs_logerror)
0.08
0.06
0.04
0.02
0.00
Basement
NaN
storydesc









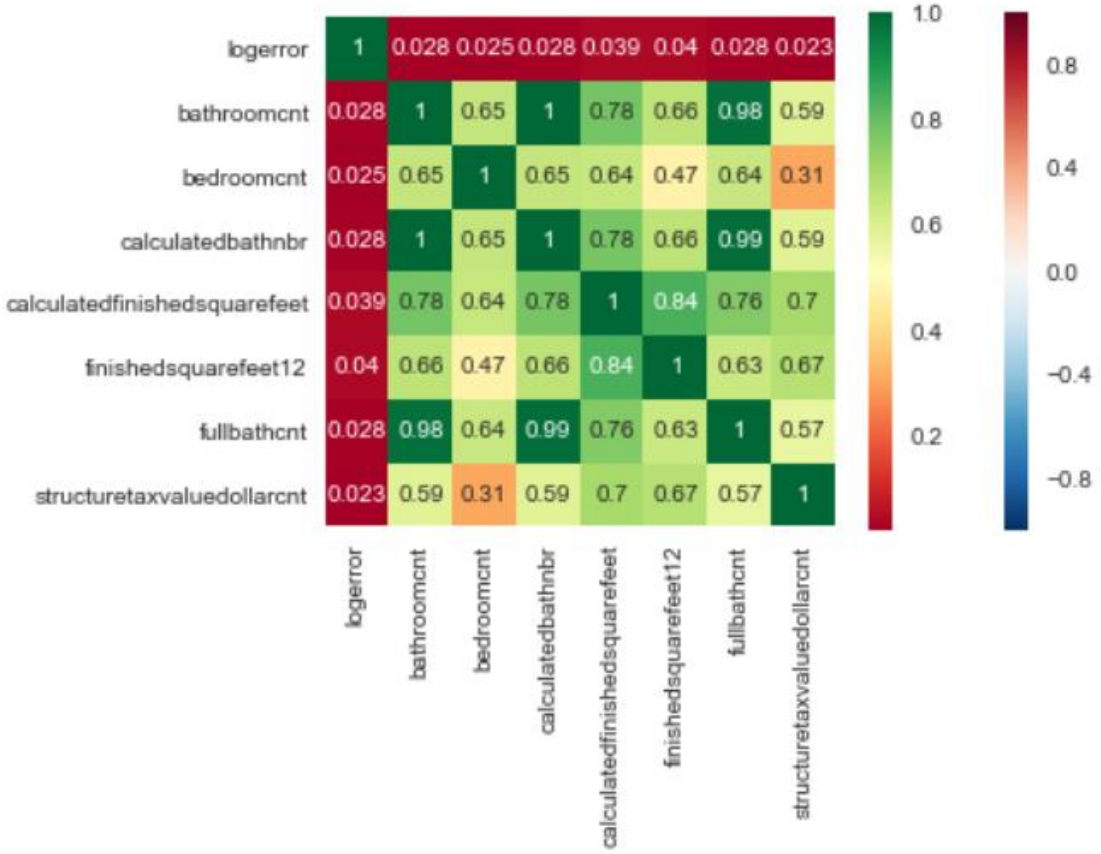


Before jumping to conclusion of the Explanatory Data Analysis, let's try to remove the features which have more than 50% of values missing in them. After removal of such variables let's build correlation matrix again and have an idea how the picture change or how the list of top predictors change. PFB the updated correlation matrix, after the removal of the features with the highest rates of missing values

Correlation with the Logerror		Correlation with Absolute Logerror	
bathroomcnt	0.027889	regionidcounty	0.061964
bedroomcnt	0.025467	taxamount	0.048277
rawcensustractandblock	0.008376	taxvaluedollarcnt	0.038343
fips	0.008363	latitude	0.019167
taxvaluedollarcnt	0.006508	regionidzip	0.002952
roomcnt	0.005760	bedroomcnt	0.001220
latitude	0.004915	bathroomcnt	-0.006541
parcelid	0.004837	transaction_month	-0.008101
propertylandusetypeid	0.001003	roomcnt	-0.035022
regionidcounty	0.000341	longitude	-0.036256

After looking at the above correlation table its quite evident that the missing values are badly impacting the purpose of generalizing the relationship between logerror and independent variables. Further the correlation even gets stronger with the absolute log error.

Next step is to check the aspect of **multicollinearity,** which can be easily checked through a correlation matrix among the independent variables. PFB the correlation matrix with color coding



After looking at the above multicollinearity matrix, its quite visible that few of the independent features are highly correlated with each other and from the pair of highly correlated features one of the feature should be removed. PFB the list of features which are highly correlated with each other

1. **fullbathcnt** is highly correlated with the **bathroomcnt** i.e. correlation is approx. 0.98
2. **calculatedbathnbr** is highly correlated with the **bathroomcnt** i.e. correlation is approx. 1
3. **fullbathcnt** is highly correlated with the **calculatedbathnbr** i.e. correlation is approx. 0.99

Summary of the Explanatory Data Analysis:

- During the outlier analysis it was observed that in logerror there seems to outliers at the both extremes of the logerror distribution.
- While exploring the transaction time there seems to be higher number of sales
    - During 6th Month of the 2016
    - During 1st,8th,15th,22nd of the very month
    - Properties sales were higher during the Friday of every week
- There seems to be high concentration of missing values in the given data set, even few of the columns have missing values in 99% of the rows.
- Location wise most of the data was situated in populated area
- Correlations of variables with log error are quite weak. However correlations become stronger with the absolute log error. But still they seems to be weak.
- During the categorical variables analysis following points was observed.
    - Airconditioned building with evaporative cooler has the highest logerror and absolute log error. Both of the log error and absolute log error were on positive and higher positive side means prices were predicted higher but actually they were lower
    - Buildings with Radiant heating system were observed with highest absolute log error. While mean logerror values were on lower positive side means predictions by Zillow machine learning algorithm are predicting higher and lower house prices wrt actual prices
    - The property which was previously used for the purpose of the Store/Office have the highest logerror and absolute logerror.
    - Properties built with the Frame infrastructure have the highest absolute log error.
    - Properties built with the highest French provincial have the highest absolute logerror.
- Few of the variables related to bath room counts seems to be highly correlated.

Algorithm & Techniques:

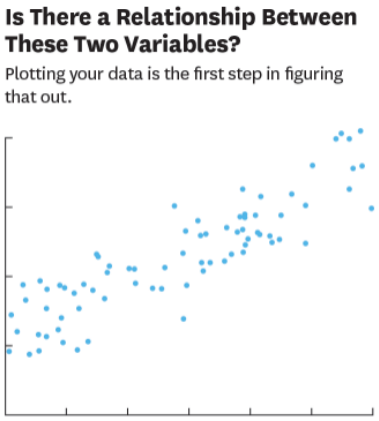
Before jumping to the details of my approach, let's have a fair idea of the algorithms being proposed in the upcoming methodology details:

How Linear Regression Works:

Regression analysis is a way of mathematically sorting out which of those variables does indeed have an impact. It answers the questions: Which factors matter most? Which can we ignore? How do those factors interact with each other? And, perhaps most importantly, how certain are we about all of these factors?

In regression analysis, those factors are called variables. You have your **dependent variable** — the main factor that you're trying to understand or predict. In Redman's example above, the dependent variable is

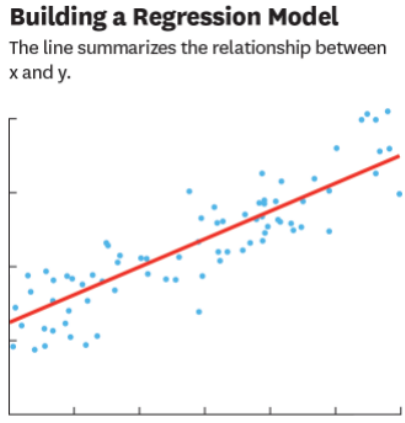
monthly sales. And then you have your **independent variables** — the factors you suspect have an impact on your dependent variable.



The y-axis is the amount of sales (the dependent variable, the thing you're interested in, is always on the y-axis) and the x-axis is the total rainfall. Each blue dot represents one month's data—how much it rained that month and how many sales you made that same month.

Glancing at this data, you probably notice that sales are higher on days when it rains a lot. That's interesting to know, but by how much? If it rains 3 inches, do you know how much you'll sell? What about if it rains 4 inches?

Now imagine drawing a line through the chart above, one that runs roughly through the middle of all the data points. This line will help you answer, with some degree of certainty, how much you typically sell when it rains a certain amount.



This is called the regression line and it's drawn (using a statistics program like SPSS or STATA or even Excel) to show the line that best fits the data. In other words, explains Redman, "The red line is the best explanation of the relationship between the independent variable and dependent variable."

In addition to drawing the line, your statistics program also outputs a formula that explains the slope of the line and looks something like this:

$$Y = 200 + 5X + \text{error term}$$

Ignore the error term for now. It refers to the fact that regression isn't perfectly precise. Just focus on the model:

$$Y = 200 + 5X$$

What this formula is telling you is that if there is no "x" then Y = 200. So, historically, when it didn't rain at all, you made an average of 200 sales and you can expect to do the same going forward assuming other variables stay the same. And in the past, for every additional inch of rain, you made an average of five more sales. "For every increment that x goes up one, y goes up by five," says Redman.

Now let's return to the error term. You might be tempted to say that rain has a big impact on sales if for every inch you get five more sales, but whether this variable is worth your attention will depend on the error term. A regression line always has an error term because, in real life, independent variables are never perfect predictors of the dependent variables. Rather the line is an estimate based on the available data. So the error term tells you how certain you can be about the formula. The larger it is, the less certain the regression line.

The above example uses only one variable to predict the factor of interest — in this case rain to predict sales. Typically you start a regression analysis wanting to understand the impact of several independent variables. So you might include not just rain but also data about a competitor's promotion. "You keep doing this until the error term is very small," says Redman. "You're trying to get the line that fits best with your data." While there can be dangers to trying to include too many variables in a regression analysis, skilled analysts can minimize those risks. And considering the impact of multiple variables at once is one of the biggest advantages of regression.

How K Means Works:

K-means clustering belongs to the non-hierarchical class of clustering algorithms. It is one of the more popular algorithms used for clustering in practice because of its simplicity and speed. It is considered to be more robust to different types of variables, is more appropriate for large datasets that are common in marketing, and is less sensitive to some customers who are outliers (in other words, extremely different from others).

For K-means clustering, the user has to specify the number of clusters required before the clustering algorithm is started. The basic algorithm for K-means clustering is as follows:

Algorithm

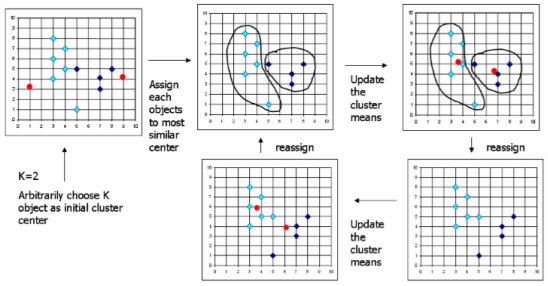
- Choose the number of clusters, k.
- Generate k random points as cluster centroids.
- Assign each point to the nearest cluster centroid.
- Recompute the new cluster centroid.
- Repeat the two previous steps until some convergence criterion is met. Usually the convergence criterion is that the assignment of customers to clusters has not changed over multiple iterations.

A cluster center is simply the average of all the points in that cluster. Its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. Consider Joe, Sam, and Sara in the

example above. Let us represent them based on their importance ratings on Premium Savings and Neighborhood Agent as: Joe = {4,7}, Sam = {3,4}, Sara = {5,3}. If we assume that they belong to the same cluster, then the center for their cluster is obtained as:

Cluster Centroid Z = (z1,z2) = {(4+3+5)/3, (7+4+3)/3}

z1 is measured as the average of the ratings of Joe, Sam, and Sara on Premium Savings. Similarly, z2 is measured as the average of their ratings on Neighborhood Agent. Below figures provides a visual representation of K-means clustering.



As discussed during capstone proposal. I will be using following methodology to solve the problem.

1. Initially I will use only the linear regression model to have an initial benchmark of the relationships
2. Then I will apply clustering algorithm like Kmeans over the clusters and then for each cluster I will apply linear regression or Classification and Regression Trees separately.

Benchmarking Technique:

I will compare accuracy of the both approaches that how much the accuracy of the second step boosted up as compared to step 1. For the comparison I will be tracking R Square in step 1

Benchmark:

To benchmark my proposed methodology I will be using the simple linear regression initially. For this the

Linear Regression:

Data Preprocessing: For linear regression I have done following preprocessing on the provided dataset

1. **Null Replacement:** First of all I have replaced Null replacement in a quick way by replacing every variable with -1.

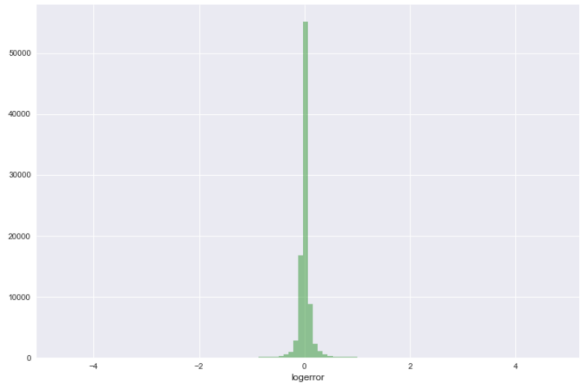
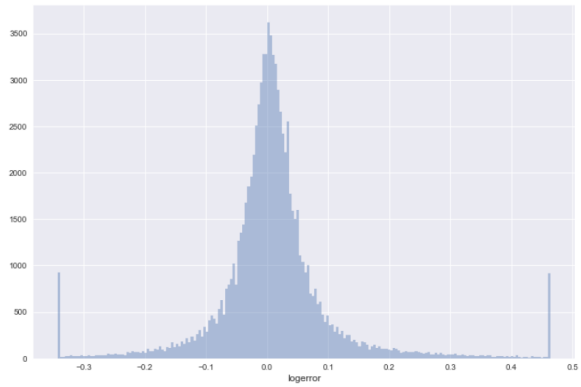
Implementation:

I am using the Sckitlearn library from the python, to implement the benchmark linear regression model. The biggest challenge I have faced was the linear regression from sckitlearn library can't be applied on columns having NaN values, so I replaced the values with -1. After removing the NaN value the multiple linear regressions was applied the benchmarking criteria of R squared resulted into a value of **0.006**. Which is quite low, **perhaps that's the reason that Zillow has announced a price money of 1.5 Million $ on the project**. As **0.006 of R Square** is quite low for a business problem to be generalized in case of the machine learning

Part 3: Methodology

Refinement:

The first refinement which I have implemented is outliers removal and then to train model again and the motivation of outliers removal came from the Explanatory Data Analysis. PFB the data distribution before and after the outlier's removal.

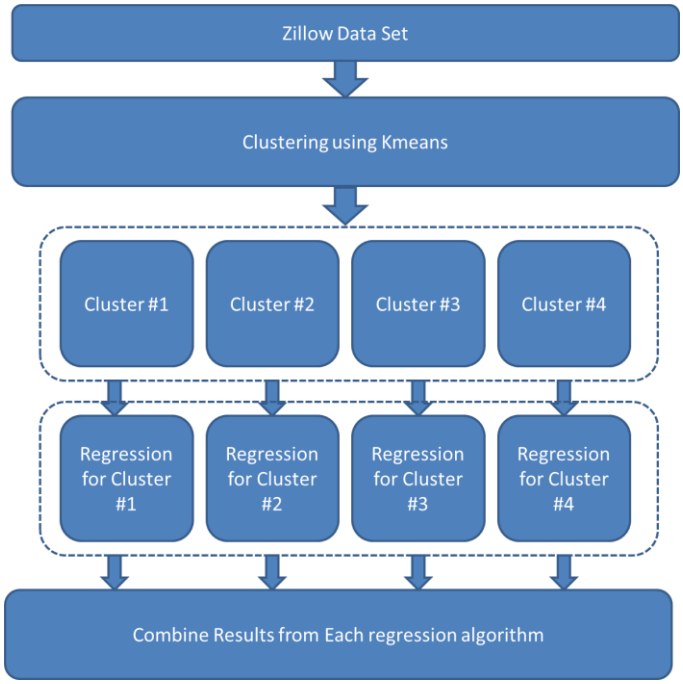
Before outliers removal (Logerror)	After outliers removal (Logerror)
 Mean_Value 0.0114572196068 Minimum_Value -4.605 Maximum_Value 4.737	 Mean_Value 0.0104040126945 Minimum_Value -0.3425 Maximum_Value 0.463882

After the outliers removal the regression model was trained again and the R Square value almost doubled. Once the regression model was trained on the data from which outliers were removed.

1. Before removing the outliers the R Squared was observed approx. **0.006**
2. After removing the outlier's, linear regression model was trained again and R Square was doubled and it reached to a value of **0.0132**. Which means outliers removal helped us to

Clustering:

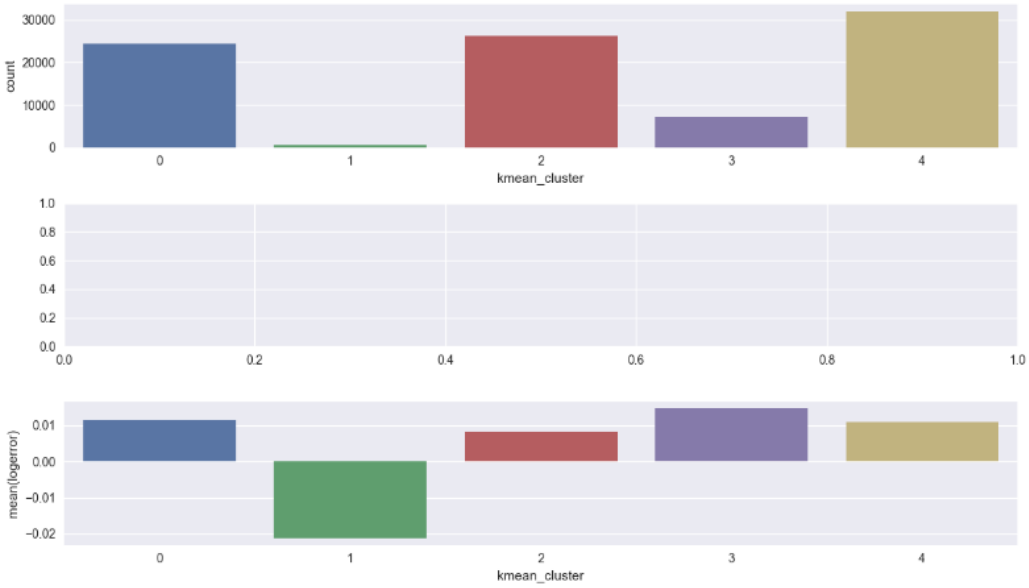
The next step in the refinement of the R Square is to apply clustering on the provided data. The refinement track will be as given below.



We will be using Kmeans algorithm from sckit learn in python to implement the K Means Clustering. Initially I would be training K means for 5,10,20,30 Clusters to have an idea that how the loge-error changes across clusters.

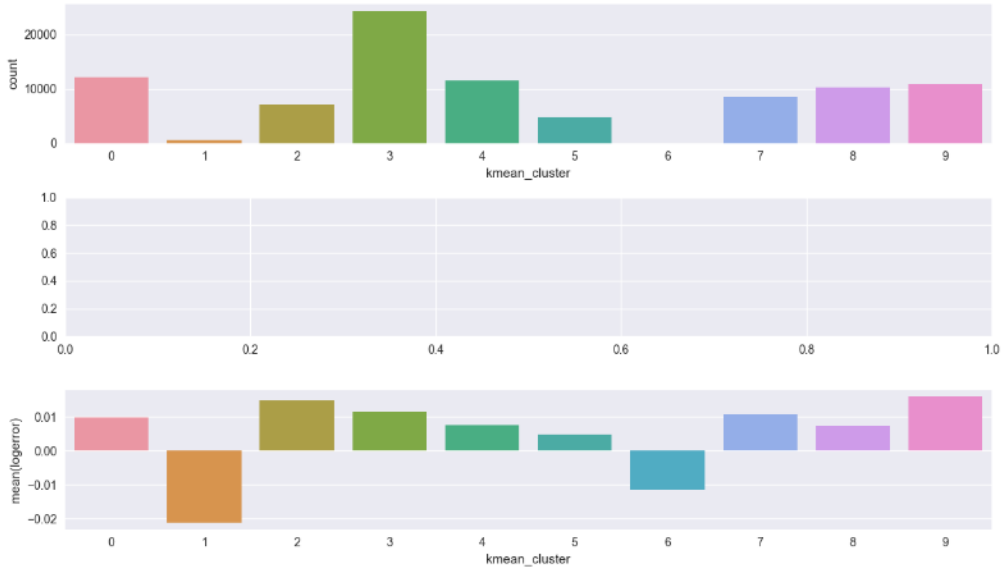
PFB the cross mapping of logerror across different KMeans Cluster output.

Five (05) Clusters:



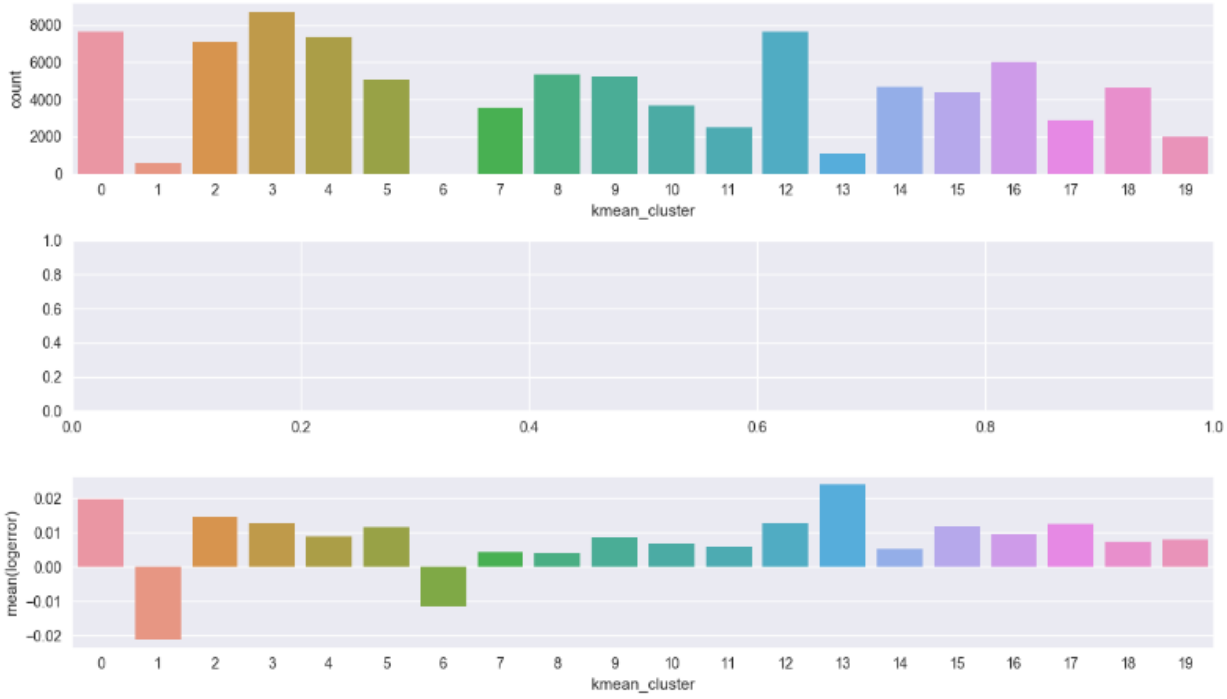
When the total of the 5 clusters were defined then against those clusters Cluster Number have average log mean error in negative but is the smallest cluster in town. While cluster number 3 have the highest average positive logerror but it's the second smallest cluster out of 5 clusters. One out of five cluster i.e. 20% of the clusters have mean error in negative.

Ten (10) Clusters:



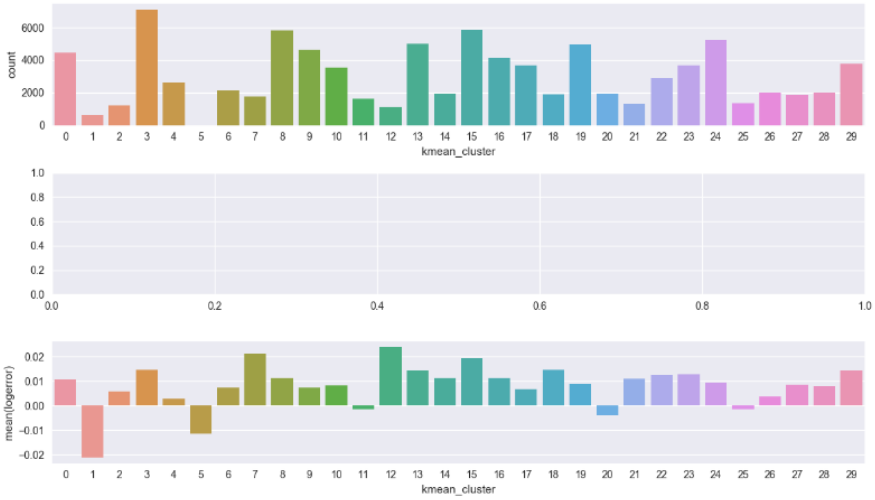
If we look at the mean logerror across 10 clusters which we got from the K means Clustering. Then again most the clusters have mean log error in positive while only 2 clusters have mean log error in negative which means 20% of the clusters have log mean error in negative.

Twenty (20) Clusters:



After observing the mean log error relationship between the clusters it was observed that only 2 clusters out of 20 clustrs have negative mean log error. Which also can be described as only 10% of the clusters have the negative mean log error?

Twenty (30) Clusters:



After looking at the 30 clusters we have observed that 5 clusters out of 30 clusters have negative mean log error. Which we can easily say that 20% of the clusters have negative mean log error.

So I have decided to move forward with the 10 clusters in hand on the Zillow data. Now I will start applying the linear regression on each cluster separately & will check how the R square increased. Benchmark for the R Square is 0.006 from we started initially

Part 4: Results

Results:

Model Evaluation and Validation

Following table of R Squared can be drawn after training separate linear regression on each Cluster.

Cluster Number	0	1	2	3	4	5	6	7	8	9
R Square	0.020374	0.15081	0.029259	0.021988	0.017574	0.024125	1	0.017725	0.026644	0.025695

It's quite visible from the above table that against few clusters the R Square increased significantly, picture can be further cleared if we can draw another table where we can have ratio between R Square over each cluster and R Square from a one regression model over complete data set. PFB the table.

Cluster Number	0	1	2	3	4	5	6	7	8	9
R Square	0.020374	0.15081	0.029259	0.021988	0.017574	0.024125	1	0.017725	0.026644	0.025695
One Regression over complete data set	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
Number of Times the R Square Increase	3.40	25.13	4.88	3.66	2.93	4.02	166.67	2.95	4.44	4.28

It's quite clear from the above table that the R Square on average increased 6 time as compared to our previous approach where only one model was being trained for complete data set, where R Square was observed around 0.006.

Further to check the generalization of the model or clustering, Hold out method was used to check the results and the results or above table results same for both training & validation dataset.

Justification

It's quite clear from the above table that regression over the clustering approach performed really well as compared to one regression model over complete dataset. The major gain which we achieved in the enhancement of the R Square was due to clustering. As building one model for whole dataset was creating problems for the regression model to generalize it well. So clustering helped us to divide dataset into multiple dataset which have similar behavior in each cluster but yet clusters behaviors was quite different as compared to other clusters. So every cluster has now similar behavior in each cluster but the different as compared to rest of the clusters, so regression can be easily generalized over a one particular behavior as compared to a mix of behavior where we received the R Square score of 0.006 which was quite low.

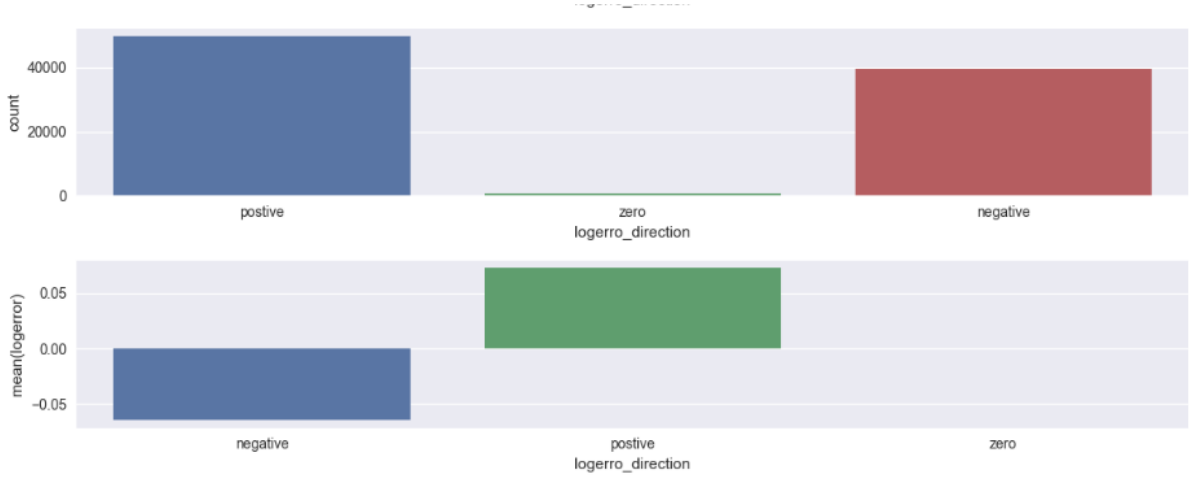
Part 5:

Conclusion

Visualization

In the free form visualization section of the project report I will try to explore whether the exploration of +ve log error, zero log error or negative log error is easy. For this I shall create a derive feature from the logerror feature and then I will train a decision tree over it and further will take help of the confusion matrix/accuracy to gauge whether positive log error can be easily predicted or negative log error can be easily predicted.

Let's have a look on the categorization of the log error in three categories and it seems that there are more values with the positive log errors than negative log error. While there are only few with zero log error.



After training a Decision Tree from the sckitlearn library and checking the accuracy of the each class in the data set i.e. positive_logerro, negative_logerr & zero_logerror. Accuracy of the negative came higher as compared to other classes i.e. 21%. While the lowest accuracy was observed against zero_logerror, which was around 4%. Which means predicting the zero logerror is the difficult on portion of the data provide.

Reflection

I have selected the Zillow log error prediction as part of the Machine Learning Nano degree capstone project. During the implementation of my project I have followed following steps to reach to any conclusion

1. **Initially Training File features were investigated separately**
2. **Then Features of the properties files were investigated separately**
3. **Then we started the exploratory data analysis of the training file where we observed**
    a. **Outliers contribution in the log-error**
    b. **Month, Date, Day of the week distribution against the property sold.**
4. **Then we started the explanatory data analysis over the properties file, where we explored the properties files from the following aspects**
    a. **Initially we explored the missing values and missing values ratios were observed quite high**
    b. **Then we plotted the correlation matrix and checked which of the independent variable have strong correlation with the dependent variable**
    c. **Further to have a confidence over the insights from the correlation matrix, matrix plots were drawn**
    d. **In next step we analyzed the categorical variables and their behavior with the dependent variable i.e. logerror**
    e. **Then we removed the features which had the missing values greater than 50 & updated correlation matrix was built where the stronger correlations were emerged.**
    f. **In the final step of EDA we removed collinear independent variables**
5. **In the next step for the model building following steps were followed**
    a. **Initially we trained linear regression and R Square was observed quite low**
    b. **In the next step we removed outliers and trained linear regression, this time the R Square was increased again**
    c. **In the next step K Means clustering was utilized and K Means algorithm was trained for 5,10,20,30 clusters. And model trained against 10 clusters was finalized.**

d. **Then 10 different linear regression algorithms were trained against each cluster and we observed a significant improvement in the R Square at the end.**

During the complete machine learning implementation framework for Zillow logerror prediction. The most interesting part which I have observed during the project was that provided features in the dataset have strong correlation with the Absolute Logerror as compared to absolute error. Which clearly describes that it would be difficult to predict that whether the actual price of the house would be above the predicted price or would be lower as compared to predicted price? But based on the provided data it's easy to predict whether the predicted price would have the difference with the actual price or not.

Improvement

As per my understanding that missing values in the Zillow dataset is the biggest challenge as high ratio of missing values represent that data is quite sparse and most of features are badly impacted by missing values. And that was reason (high missing value) that Zillow put a prize worth 1.5 Million $ prize over the problem solving using the machine learning algorithm. Following proposed strategy can be used for better null replacement.

Proposed null replacement strategy should be location based, since data for the location is provided in the provided dataset. And property/rea estate is always a location sensitive as the location of property describes a lot of information like if it's situated in the commercial area so higher prices should be expected, if the property is park facing higher prices should be expected etc. So For null replacing we should use geo fencing i.e. for missing value replacement the average feature value within 10-20 kilometers should be selected. For example if the number of the rooms isn't mentioned in the Zillow data than missing value should be replaced with the average number of rooms in the properties within 20 Kilometer circle. As such approximation would give a fair idea of the room number in the subject property for which the number of rooms was missing.

Second proposed improvement can be addition of derived variables. i.e. induce new variables derived from existing variables. For example from the transaction date we can extract new feature like Month of the year, Day of the month, week of the year etc. As such variables can also provide us some sort of information regarding log error.

Third & the last suggestion to improve the results could be to use the regression models with polynomial degree. i.e. I have used the linear regression for the subject project but we can use the regression model with degree of 2 or 3 or n. Which will help us to better fit the data and generalize it well & ultimately R Square will be enhanced significantly.