Data Science Course Project

Wenyuan Lu

University of Pittsburgh
Course: BMIS 2542 Programming Essentials for Data Science
Professor: Narayan Ramasubbu

Data Science Course Project: Zillow's Home Value Prediction

1 Introduction

1.1 Problem Definition

This project is going to focus on the housing market and prices of houses which based on a competition started by Zillow.com, a real estate Web site launched in February 2006 and posted on Kaggle.com with the help of Python (libraries including Sklearn, panda, Numpy) and data visualization tool Tableau. From the competition, the task was to predict differences between the 'Zestimate' (the estimation made by Zillow.com) and the actual sale price.

Generally, estimating the price of houses is a tough question and unlike stock price or weather forecasts which can be validated every day, even every second and adjust the forecasting model immediately, the prices of houses can only be validated when the transaction happens but it may take dozens of years for one particular house to be sold and validate the price estimate. For example, when we are trying to validate the estimate of price about one house, it is impossible that this house happens to be sold, therefore, it is hard to adjust the model.

In history, lots of researchers were trying to focus on repeated transactions, or geographically distribution to generate the mathematics models to better estimate the housing prices. However, with the emergence of big data and company doing the housing price estimation like Zillow.com, it is possible to generate a big data perspective research and help to improve the housing market. For Zillow.com, there are other methods to improve the accuracy of the estimation, one possible way is to analysis the photos uploaded by the home owners with deep learning image recognition technology.

1.2 Project Background

The reason why I am interested in this topic is firstly, this problem is a real-life problem and the data is real-life data instead of fake data which means that this project is relatively meaningful because the solution may help to reorganize the housing market and help both the sellers and buyers to understand the true price of the houses. Over years, estimations of home values have been a considerable important. From homeowners, neighbors, government tax assessors, to appraisers working for real estate agents have been trying to establish mathematics models to estimate precise values based on the available market information. The construction business is obviously vital as far as monetary activity, not just on account of its commitment to Gross Value Added, yet in addition in view of the way that it powers whatever remains of monetary parts and impacts business. Additionally, inside the construction business, the estimation has uncommon attributes due to the social effect it involves. A house is a risk of most extreme significance and getting one request an incredible monetary exertion (Epley, 2016).

Secondly, this dataset is uploaded by Zillow itself so the dataset is relatively reliable. Meanwhile, the housing transactions details are public in America which implies that the data analysis will be legal and reasonable.

Finally, my family has been using Zillow for our first-hand data collecting method since we were trying to buy a new house at the beginning of this year. However, we sometimes found some obvious estimation errors that Zillow.com make and sometimes we were confused, so I want to apply the data analysis techniques I learned from this course to find why Zillow may make the mistakes and how to use the estimates made by Zillow.com better.

1.3 Project Target

To summarize, this project is going to discuss:

1) Analyze the general patterns of the housing market from the housing details to see geographic distributions (take Southern California as an example) while cleaning and exploring data clearly.

2) Analyze the patterns of the difference between the Zestimate and the true sale price such as how accurate are Zillow doing now? Are they improving their models from time to time? Do they tend to overestimate the price or underestimate the price?

3) Try to use the supervised models including linear regression model with Python "sklean" library to predict the differences between the Z-estimate and the true price when transaction happens.

2 Literature Review

2.1 Overview of Zillow

Zillow.com is a real estate web site launched in February 2006. Zillow provides an estimate of market value for over 46 million homes based on a proprietary formula. In general, it offers free value estimates, or the so-called 'Z-estimates', which is going to be a significant feature in this project, using data from appraisal districts and from multiple listing services (MLSs), depending on availability. Zillow currently also accepts for-sale listings, offers information about buying and selling, and provides links to mortgage providers and real estate professionals. Several of these latest information offerings have been added since the data for this study was collected (Zillow.com).

In general, The Z-estimate home value is Zillow's estimated market value for an individual home and is calculated for about 100 million homes nationwide. It is a starting point in determining a home's value and is not an official appraisal. The Zestimate is automatically computed daily based on millions of public and user-submitted data points.

2.2 Z-estimate Introduction

The Z-estimate Forecast Methodology can be divided into several stages. Firstly, to forecast a path for the individual Zestimate, Zillow rely on two different types of data. The first is the general level of Zillow Home Value Forecast (mainly county level) which forecasts the Zillow Home Value Index (ZHVI) and is produced using a variety of economic and housing data. The forecast is then combined with data computed by detailed characteristics of the property, including its features like numbers of bedrooms and numbers of bathrooms as well as the past behavior of its Z-estimate which may help to adjust the prediction model in a great extent. The forecasting stages will eventually concentrate on how those aggregate forecasts, in combination with property characteristics, are used to construct the forecast for a particular property.

After computing the predicted Zestimate in 12 months for a property, the path for the Z-estimate will be smoothly connects the forecasted value to the current value of the Z-estimate. Here, an important method against over-fitting is shrinking towards the forecast for the county as a whole, with the shrinkage weights gradually declining over time. Currently, as posted on Zillow.com, the Z-estimate forecast is available for more than 50 million individual properties spanning 550 counties across the entire country

based on their data gathered since 2006 when Zillow.com was founded. The predictive accuracy of these one-year forecasts was assessed by back-testing the model over the past five years. Back-testing consists of running forecasts on historical. The below table summarizes the average absolute percentage error of the 12-month forecast.

From the figure 1, it can be noticed when comparing to a naive forecast based on a simple random walk model and to simply extending county forecast to individual properties, the construction of a property-specific forecast significantly improves accuracy.

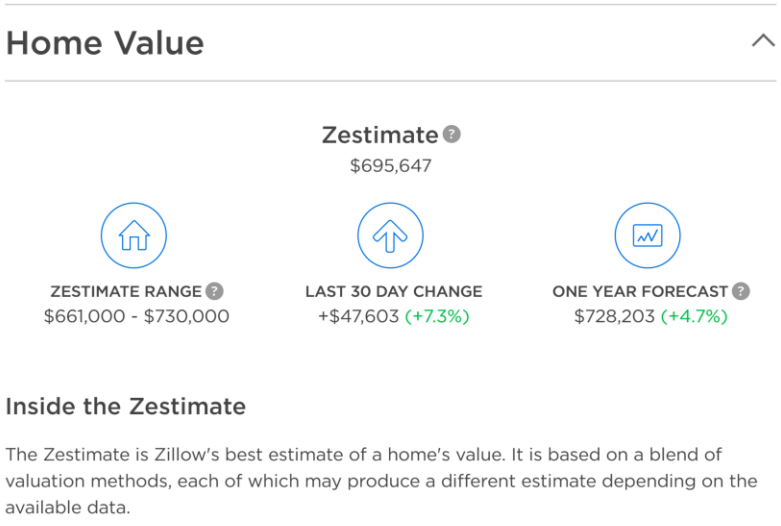
Figure 1 Naïve Forecast, County Forecast, Zestimate Forecast Comparison

Model	Average Absolute %	Improvement over Naïve
Naïve Forecast	7.35%	0%
County Forecast	6.47%	11.9%
Zestimate Forecast	5.84%	20.5%

Source: Zillow.com, https://www.zillow.com/zestimate/

Take Figure 2 as an example, this figure is a screen cut from Zillow.com and indicates a house's price. The value range, which is indicated in Zillow.com and related to the Zestimate, shows the high and low estimated values of a home. Therefore, less historic data may lead to a wider, vice versa, a smaller value range means lots of information are available to compute the Zestimate and Value Range. For the statistically minded, the Value Range is a 70 percent confidence interval which means Zillow is 70 percent confident that the true house price is in the value range.

Figure 2 screen cut from Zillow.com



Source: Zillow.com, https://www.zillow.com/homes/for_sale/Pittsburgh-PA-15213/11407860_zpid/63944_rid/40.465527,-79.9188,40.419932,-79.992614_rect/13_zm/

Zillow use proprietary automated valuation models that apply advanced algorithms to analyze the data to identify relationships within a specific geographic area, between the home-related data and actual sales prices. The housing characteristics are given different weights according to their influence on home sale prices in each specific geography over a specific period of time, resulting in a set of valuation rules, or models that are applied to generate each home's Zestimate. To conclude, Zillow uses aggression models and give different weights to all attributes. Currently, it is reported that Zillow has data on 110 million homes and Z-estimates and Rent Z-estimates on approximately 100 million U.S. homes (Zillow Internal, 2013). With the advanced algorithms and the machine learning techniques, the Z-estimated are calculated automatedly by software models designed by statisticians. From the customer perspective (including potential buyers and sellers), the most significant question regarding Zillow's Z-estimates is still whether they reflect transaction prices. According to Neiva (2017), Zillow has been described both as "a useful site" and as “categorically wrong." Many instances of complaints from homeowners using Zillow to estimate the value of their homes are heard though they understand that Zillow’s Z-estimate may have flaws but they are still extremely upset when they found that Zillow underestimate the prices of their houses.

Therefore, the objective of this project is to examine the relationship between Zillow's Z-estimates and actual transaction prices, while also examining where Zillow model may face regular mistakes.

3 Data Collection and Procedures

3.1 Data Source

As mentioned above, this data set is uploaded by Zillow.com onto Kaggle.com as the dataset for the competition. This dataset is separated to several parts. First one is the details of houses which have about 60 features including the longitude, latitude, Number of bathrooms in home including fractional bathrooms, Number of bedrooms in home, Number of pools on the lot (if any) to type of cooling system present in the home (if any). The second data file includes transaction date and the target, the so-called ‘Log-Error’. As mentioned above, people face problems when they are trying to validate the reliability of housing price estimation models. Therefore, Zillow come up with a smart way to validate the preciseness of the predictions by setting a variable called log error, and this log error is defined as equation (1):

$$\text{logerror} = \log(\text{Zestimate}) - \log(\text{SalePrice}) \quad (1)$$

The idea of log error is brilliant. As mentioned above, generally, estimating the price of houses is a tough question and unlike stock price or weather forecasts which can be validated every day, the prices of houses can only be validated when the transaction happens but it may take dozens of years for one particular house to be sold and validate the price estimate. For example, when we are trying to validate the estimate of price about one house, it is impossible that this house happens to be sold, therefore, it is hard to adjust the model. The calculation of log errors help to solve this problem, by calculating log errors, the difference between the Z-estimate and the true sale price can be recorded when the transaction happens.

This log error is recorded in the transactions training data. By analyzing the Log-Error, we may able to find the relations between the Z-estimate and the sales prices, or rather, the accuracy of the Z-estimate and eventually improve the model. Finally, because my family lives in Los Angeles, so I am

more interested in the housing market of Southern California. Therefore, I am going to take Southern California data to analyze.

3.2 Data Exploration

3.2.1 Overview of Data

The first dataset having the details of houses contains around sixty features.

Figure 3 Descriptions of Features

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 w
'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.wikipe
'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
'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
'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 assignr
'censustractandblock'	Census tract and block ID combined - also contains blockgroup assignr
'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, spl
'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, e
'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

Figure 3 shows all the description of each feature name. These features can be divided into several segments: Identification: Parcelid, to identify one specific house Physical attributes: Location, lot size, square footage, number of bedrooms and bathrooms and many other details. Tax assessments: Property tax information, actual property taxes paid, exceptions to tax assessments and other information provided in the tax assessors' records.

By inputting into DataFrame with python, I found that the shape of the DataFrame is (2985217, 58) and with the describing method, I can get table and get the overview of the data from count, mean to max.

Figure 4 Describing Dataset

	parcelid	airconditioningtypeid	architecturalstyletypeid	basementsqft	bathroomcnt
count	2.99E+06	811519	6061	1628	2.97E+06
mean	1.33E+07	1.931166	7.202607	646.883292	2.21E+00
std	7.91E+06	3.148587	2.43629	538.793473	1.08E+00
min	1.07E+07	1	2	20	0.00E+00
25%	1.16E+07	1	7	272	2.00E+00
50%	1.25E+07	1	7	534	2.00E+00
75%	1.41E+07	1	7	847.25	3.00E+00
max	1.70E+08	13	27	8516	2.00E+01
	bedroomcnt	**buildingclasstypeid**	**buildingqualitytypeid**	**calculatedbathnbr**	**decktypeid**
count	2.97E+06	12629	1.94E+06	2.86E+06	17096
mean	3.09E+00	3.725948	5.78E+00	2.30E+00	66
std	1.28E+00	0.5017	1.81E+00	1.00E+00	0
min	0.00E+00	1	1.00E+00	1.00E+00	66
25%	2.00E+00	3	4.00E+00	2.00E+00	66
50%	3.00E+00	4	7.00E+00	2.00E+00	66
75%	4.00E+00	4	7.00E+00	3.00E+00	66
max	2.00E+01	5	1.20E+01	2.00E+01	66
	yardbuildingsqft26	**yearbuilt**	**numberofstories**	**structuretaxvaluedollarcnt**	**taxvaluedollarcnt**
count	2647	2.93E+06	682069	2.93E+06	2.94E+06
mean	278.296562	1.96E+03	1.401464	1.71E+05	4.20E+05
std	369.731508	2.34E+01	0.539076	4.02E+05	7.26E+05
min	10	1.80E+03	1	1.00E+00	1.00E+00
25%	96	1.95E+03	1	7.48E+04	1.80E+05
50%	168	1.96E+03	1	1.23E+05	3.06E+05
75%	320	1.98E+03	2	1.97E+05	4.88E+05
max	6141	2.02E+03	41	2.51E+08	2.83E+08
	assessmentyear	**landtaxvaluedollarcnt**	**taxamount**	**taxdelinquencyyear**	**censustractandblock**
count	2.97E+06	2.92E+06	2.95E+06	56464	2.91E+06
mean	2.01E+03	2.52E+05	5.38E+03	13.892409	6.05E+13
std	3.68E-02	4.45E+05	9.18E+03	2.581006	3.25E+11
min	2.00E+03	1.00E+00	1.34E+00	0	-1.00E+00
25%	2.02E+03	7.48E+04	2.46E+03	14	6.04E+13
50%	2.02E+03	1.67E+05	3.99E+03	14	6.04E+13
75%	2.02E+03	3.07E+05	6.20E+03	15	6.06E+13
max	2.02E+03	9.02E+07	3.46E+06	99	4.83E+14

However, when looking at some attributes, some unexpected things may happen, for example, the Parcelid should be a string while here DataFrame consider Parcelid as a numerical object. Therefore, the

types of the features should be checked. Take first several features as examples, I saw that the type of parcelid is automatically set to int64 instead of string what I wanted and this may cause confusion in the following analysis, therefore, I use 'astype' to change the type of this feature and then the type of parcelid changes to 'object'.

Figure 5 Types of Data

parcelid	int64
airconditioningtypeid	float64
architecturalstyletypeid	float64
basementsqft	float64
bathroomcnt	float64
bedroomcnt	float64
buildingclasstypeid	float64
buildingqualitytypeid	float64

By analyzing the data generated from python, we can find a lot of information, for example, the mean value of the year built is 1960 and about 25% of houses in this dataset were built before 1950s and another 25% of houses in this dataset were built after 1980s. Since the shape of the dataset is (2985217, 58), which means it has 2985217 rows and 58 columns and this may far exceed the calculation capacity of one computer, therefore, I inputted the other dataset.

The other dataset contains 90275 rows of information including the parcelid, logerror, transaction date. By describing the dataset, the following figure can be generated. Similarly, the parcelid should be factorized.

Figure 6 Second Dataset Description

	logerror	month	day	year
count	90275	90275	90275	90275
mean	0.011457	5.849848	16.33951	2016
std	0.161079	2.81269	9.008589	0
min	-4.605	1	1	2016
25%	-0.0253	4	8	2016
50%	0.006	6	16	2016
75%	0.0392	8	24	2016
max	4.737	12	31	2016

From figure 6, it can be found that there are totally 90275 rows of records and the mean value of logerror is above 0 while 50 % of records are in the range from -4.605 to 0.006 which is above 0 which means that it can not be clarified to say whether Zillow tends to overestimate or underestimate the value of houses and I need more stat analysis. From the Zillow's accuracy's perspective, 50% of records are in the range of -0.0253 to 0.0392, which is satisfying for most potential 'customers'.

Figure 7 Bar chart of Log Error

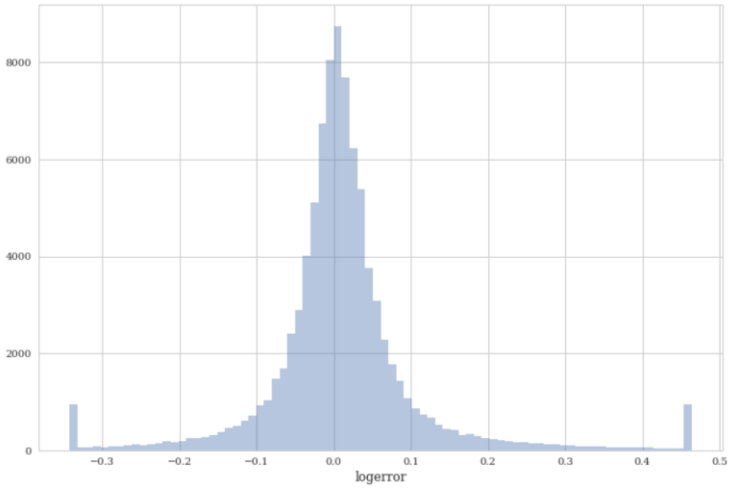
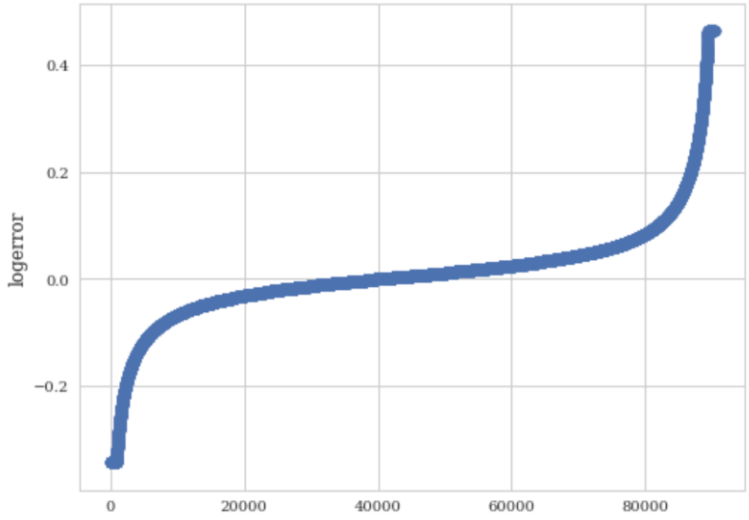


Figure 8 Cumulative Distribution of Log Error



From these two figures plotting the log error, it is obvious that the log error distribution is a beautiful normal distribution. By set a upper limit and a lower limit of 99% to the log error to give a better visualization and this is why at the two sides of the first figure, two peaks can be discovered, this does not mean that there are hundreds of log errors are at 0.45, this means that one 0.45 is the 99% limit, or rather, one percent of log errors is above 0.45 and 98% of log errors are in the range of -0.35 to 0.45. Meanwhile, the center peak which is above 8000 shows that there are more than 8000 log errors are very close to 0. The second figure is the cumulative distribution of log errors and from this figure, we can also figure out that the most of the horrors are around 0.

With the help of join method which is called merge, I am able to merge these two datasets and get a dataset which is going to be analyzed in more details. The shape of this new dataset is (90275, 63).

3.2.2 Geographic Visualization

With the longitude and latitude values, I am able to draw the map to better visualize the distribution of data From Figure Geographically Distribution of Data, it can be easily observed that the data set, or rather, the houses going to be analyzed are all located in southern California, which are located around Los Angeles county and Orange county specifically. It can be observed that the houses are distributed averagely from the color showing the intensity of houses. A randomly distributed dataset may seem to be more reliable.

Figure 9

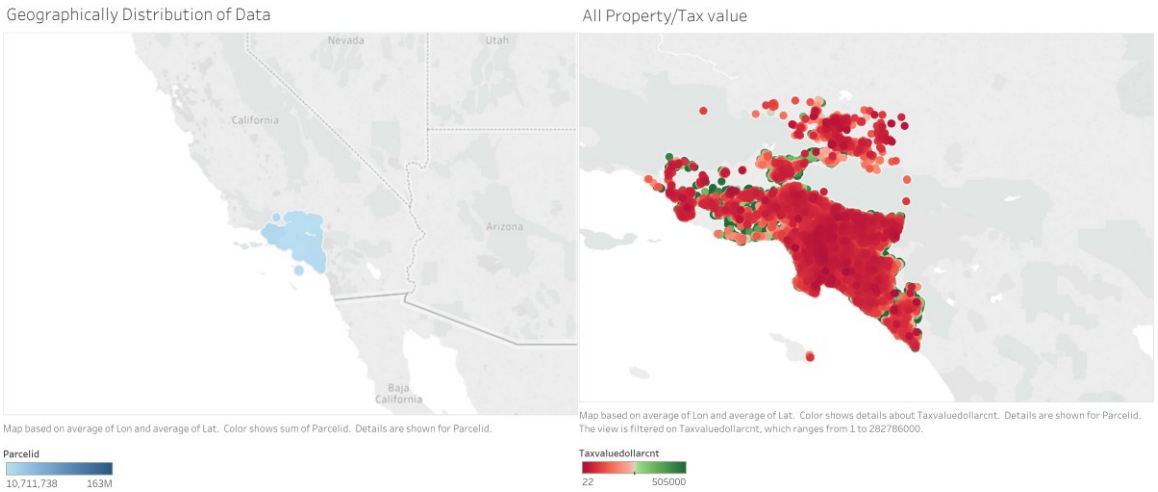


Figure 10

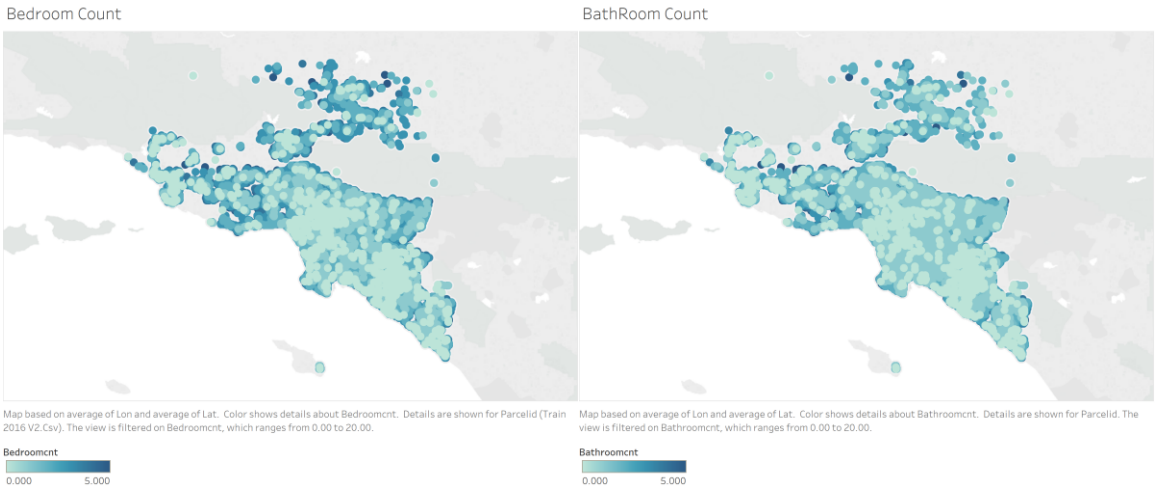
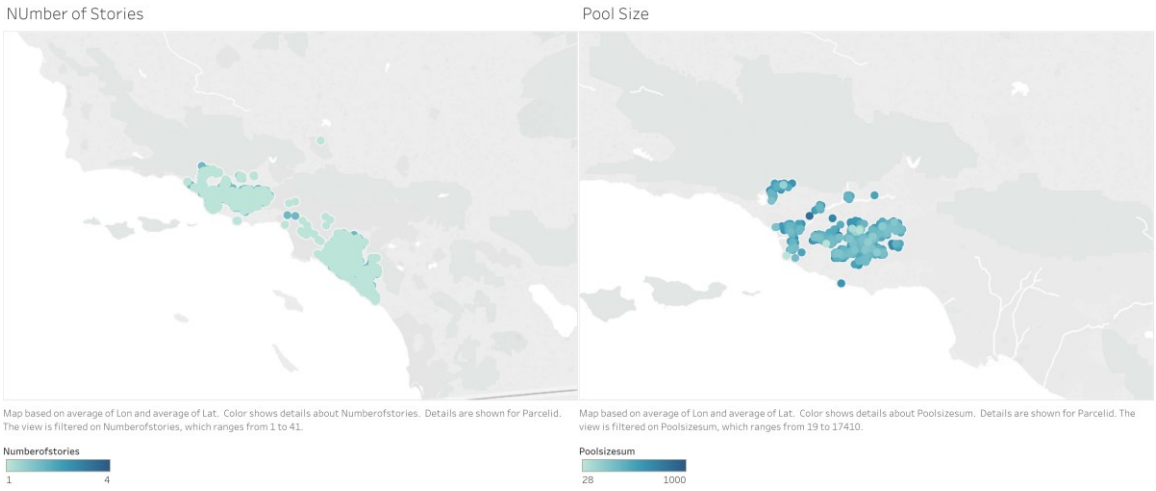


Figure 11



Tax value is a relatively important feature applying common sense, tax value is based on the tax assessment from government. Generally, From Figure All Property/Tax Value, we can recognize that some houses at the borders may appear to have higher tax value amount and it is obvious that most houses are red which means that the tax value amount is under the median number of the tax value amount range.

From Figure bedroom count, bathroom count, number of stories, pool size, it can be found that the houses located out of dense area, which may be not the center of cities are more likely to have more bedrooms, bathrooms, and bigger pools.

Figure 12

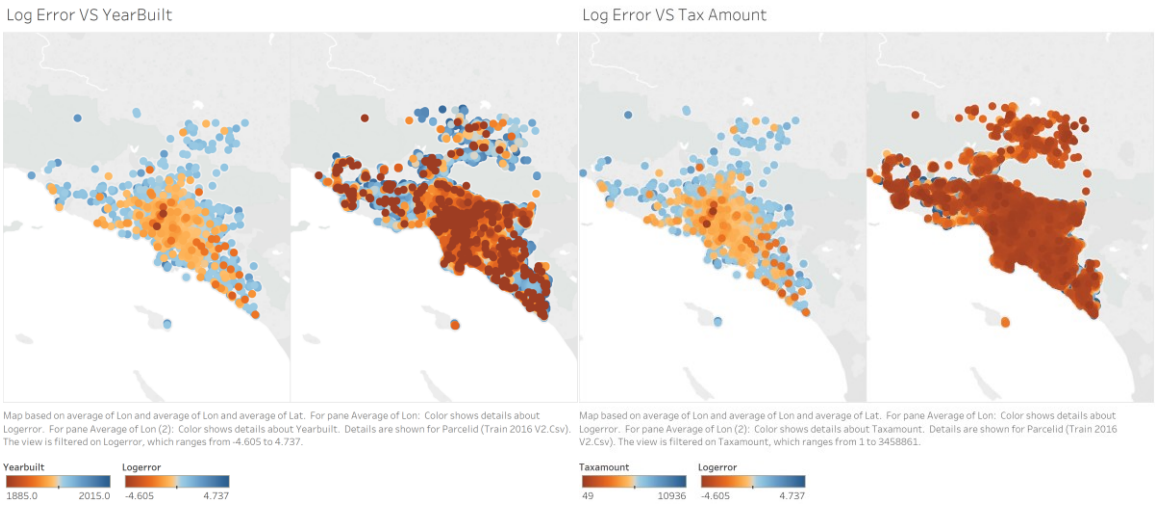
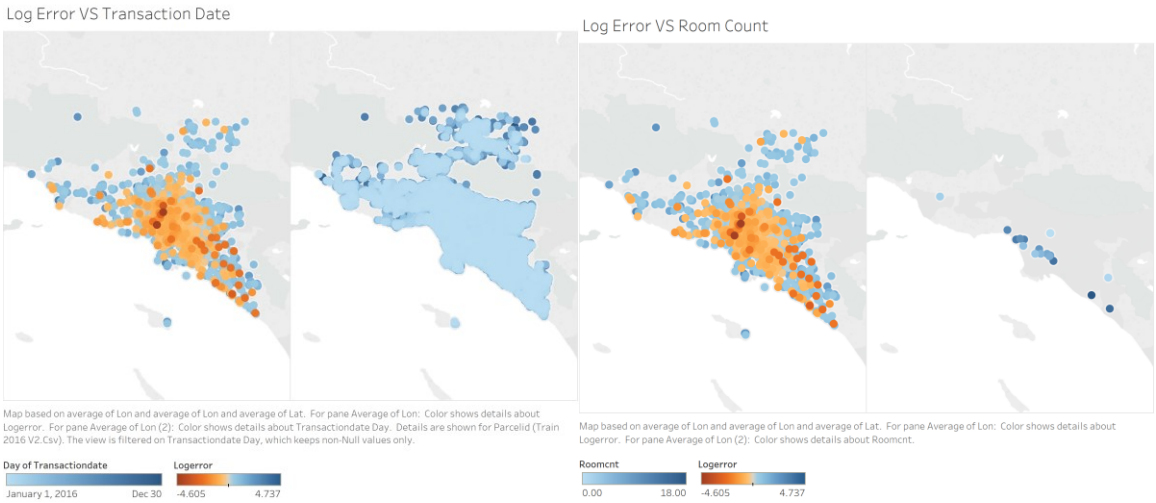


Figure 13



By plotting the most important feature: Log Error, and compare it to building year, tax amount, transaction days and room amount, I find that firstly, from the Log Error distribution, most houses are underestimated in the center of the graph because the houses located at the center have the color of red and orange to red means that the Log error is lower than zero and according to equation (1), Log error lower than zero means that Zillow underestimates the value of the house, vice versa, houses are more likely to be overestimated at the borders.

When looking at the year building, tax amount, it can be found that the data distribution pattern is similar to Log Error graph, the orange-red colors are aggerated in the center of the graph and lighter and blue colors are located at the borders. One assumption is that the prices newer houses are more difficult to estimate because of the lack of some historical data and another assumption is that the community-level estimations of these communities may be flawed and needs adjustment. According to Zillow.com, the number of transactions in a geographic area affects how much the model know about prevailing market values of homes in that area which indicated that more transactions provide more data and improve the accuracy of the Zestimate. Though Zillow.com is reported to use public as well as user-provided data for house attributes, and some areas report more data than others. The more attributes they know about homes in an area, the better the Zestimate. The estimating method used by Zillow differs from the methods concluded in the literature review including comparative market analysis (CMA). Geographically, the dataset is much larger than what old methods may think of.

3.2.3 Data Clean

From the above geographically figures, we may notice that some data are ‘invisible’ in the figure for some reason and because of the vanishing data, we are hardly to gather useful or reliable information and this happens because of the existence of null data, therefore, the null data should be cleaned. The reason for lots of null data is because most upgrade information is not in the public records, and is not easily quantifiable. The home updates and remodels information are inaccessible unless they have been reported to the local tax assessor, so those items are not included in Z-estimate calculations.

Firstly, the following figure shows the missing count and missing ratio of every feature and presents those features have more than 80% of null values.

Figure 14 Missing Raito Exceed 80%

	column_name	missing_count	missing_ratio
8	architecturalstyletypeid	90014	0.997109
9	basementsqft	90232	0.999524
12	buildingclasstypeid	90259	0.999823
15	decktypeid	89617	0.992711
16	finishedfloor1squarefeet	83419	0.924054
19	finishedsquarefeet13	90242	0.999634
20	finishedsquarefeet15	86711	0.960521
21	finishedsquarefeet50	83419	0.924054
22	finishedsquarefeet6	89854	0.995336
24	fireplacecnt	80668	0.893581
28	hashottuborspa	87910	0.973802
33	poolcnt	72374	0.801706
34	poolsizesum	89306	0.989266
35	pooltypeid10	89114	0.987139
36	pooltypeid2	89071	0.986663
37	pooltypeid7	73578	0.815043
47	storytypeid	90232	0.999524
48	threequarterbathnbr	78266	0.866973
49	typeconstructiontypeid	89976	0.996688
51	yardbuildingsqft17	87629	0.97069
52	yardbuildingsqft26	90180	0.998948
55	fireplaceflag	90053	0.997541
61	taxdelinquencyflag	88492	0.980249
62	taxdelinquencyyear	88492	0.980249

There's many methods to deal with the null values, here, I am going to use the mean method which means that I am going to fulfill all the null values with the mean values.

3.2.4 Correlation Analysis

To see whether the features have some correlations, we may use python to calculate the correlations and try to visualize the correlations with some plotting techniques. For better understanding the relationships between the features and the log errors, I created a new column called 'abs' which is the absolute value of log errors, by looking at the 'abs', I can recognize the relationships between the absolute values of log errors and the other features.

Figure 15 shows the correlations values (spearman ratio) between log error and other attributes and Figure 16 shows the correlations values (spearman ratio) between absolute value of and other attributes. However, no strong relations can be found from the figures. As what mentioned above, I am going to discuss and take another look of number of null values to see if number of null values will have some influence on the log errors.

Figure 15 Correlations values (spearman ratio) between log error and other Features

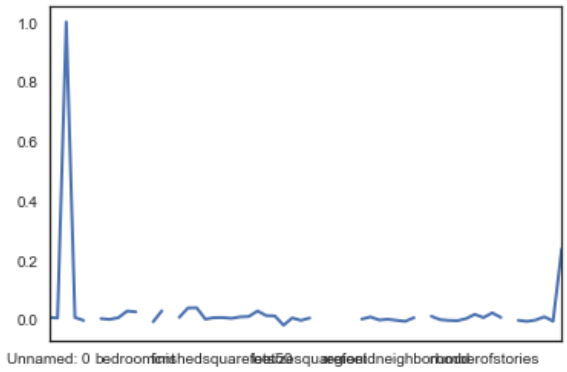


Figure 16 Correlations values (spearman ratio) between absolutes value of log error and other Features

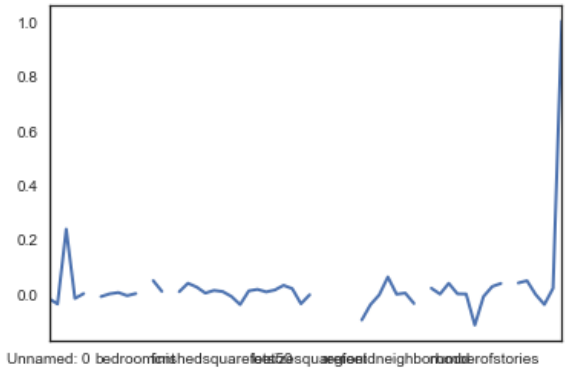


Figure 17 Heatmap of Correlations between log error and other Features

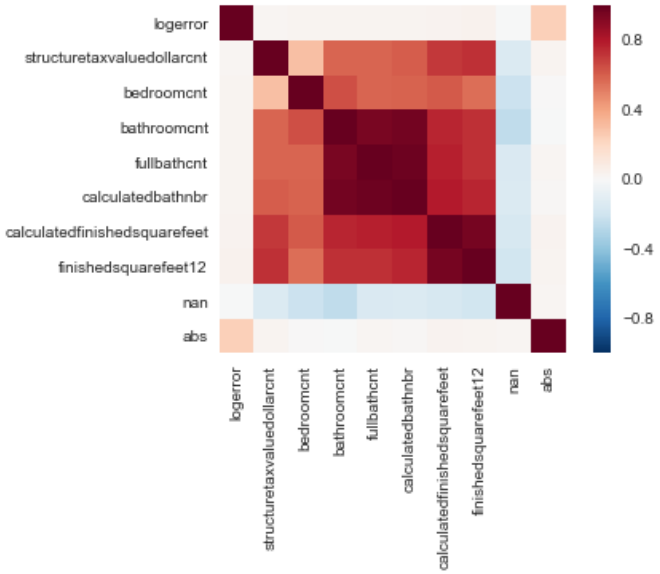
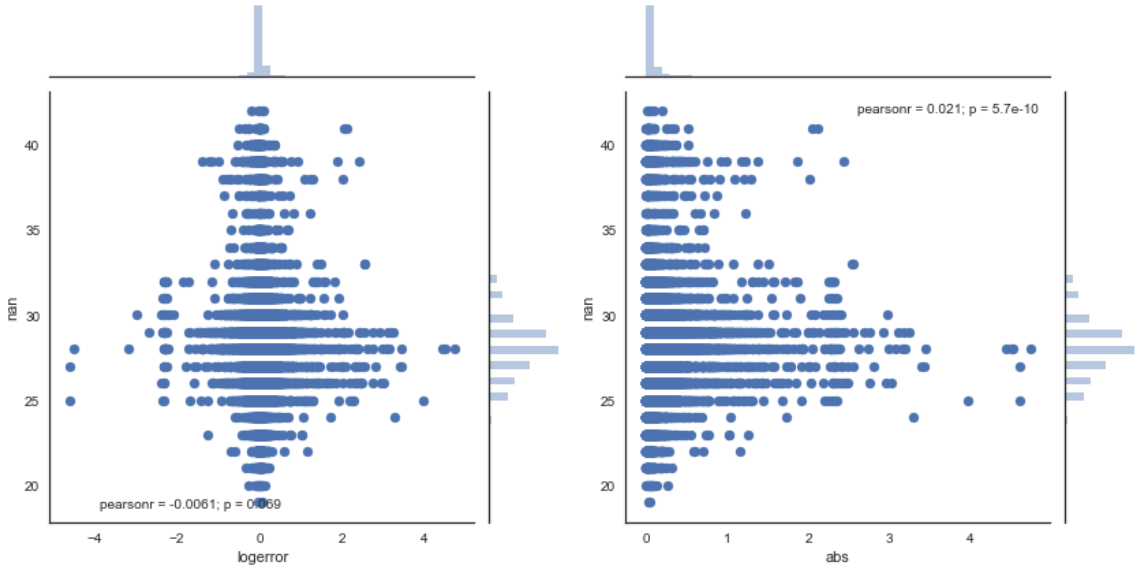
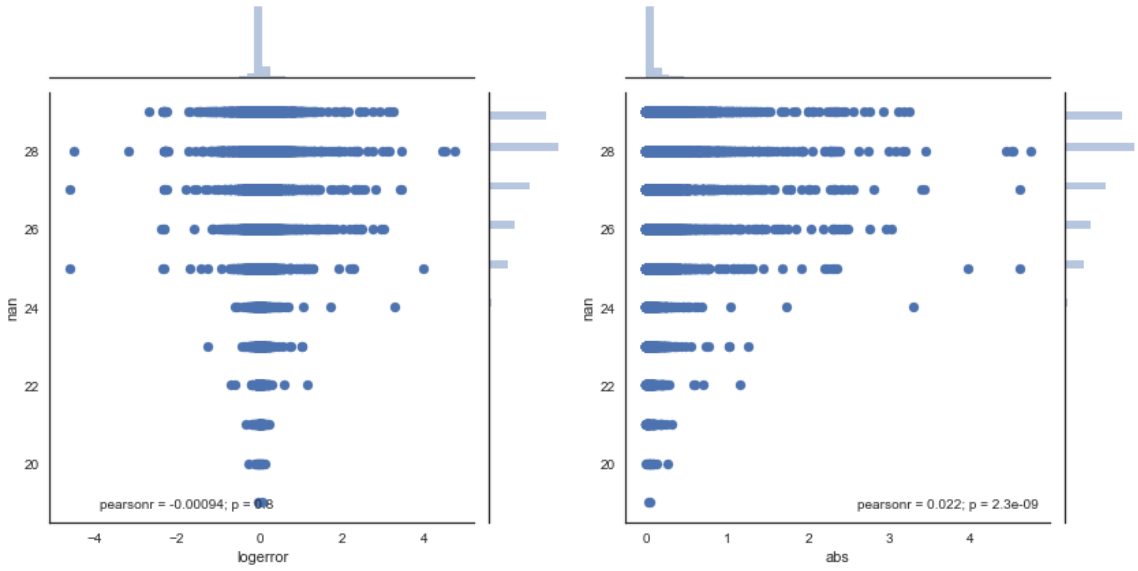


Figure 18 Log error and absolute value versus number of null values



If the assumption that number of null values do have some impacts on the predictions, then the figure should look like a great beautiful reverse triangle (which is shown in figure 19), but surprisingly, from 0 to 30 null values, the pattern seems to be reasonable, a beautiful triangle can be easily discovered. But, when the number of null values is smaller than 20, the log errors are extremely close to zero and when the number of null values increases to 25 to 30, than the log errors seems to expand quickly which is unexpected.

Figure 19 Log error and absolute value versus number of null values which are smaller than thirty



To figure out what happens to the reasons for the outliers, I made some assumptions, firstly, is this because of some missing of particular features which are very important? To validate this, I selected the records having log error larger than 0.5 (shape of (1254, 66)) and see the missing values and compare

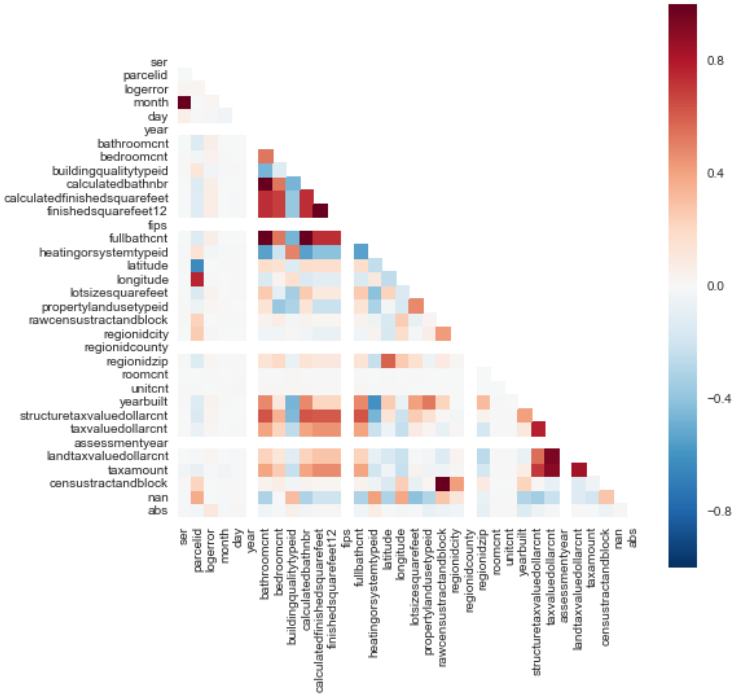
them with the missing ratio of all data, however, no strong pattern was found from this, so I suppose that this is not the reason for the extremely inaccurate log errors.

Figure 20 missing ratio of outliers versus missing ratio of entire dataset

	column_name	missing_count	missing_ratio	missing_count_all
49	typeconstructiontypeid	1253	0.999203	0.996688
19	finishedsquarefeet13	1253	0.999203	0.999634
55	fireplaceflag	1253	0.999203	0.997541
8	architecturalstyletypeid	1253	0.999203	0.997109
9	basementsqft	1252	0.998405	0.999524
47	storytypeid	1252	0.998405	0.999524
12	buildingclasstypeid	1251	0.997608	0.999823
52	yardbuildingsqft26	1250	0.996810	0.998948
34	poolsizesum	1248	0.995215	0.989266
15	decktypeid	1248	0.995215	0.992711
35	pooltypeid10	1245	0.992823	0.987139
36	pooltypeid2	1240	0.988836	0.986663
22	finishedsquarefeet6	1239	0.988038	0.995336
28	hashottuborspa	1231	0.981659	0.973802
51	yardbuildingsqft17	1226	0.977671	0.970690
62	taxdelinquencyyear	1195	0.952951	0.980249
61	taxdelinquencyflag	1195	0.952951	0.980249
21	finishedsquarefeet50	1182	0.942584	0.924054
16	finishedfloor1squarefeet	1182	0.942584	0.924054
48	threequarterbathnbr	1171	0.933812	0.866973
24	fireplacecnt	1170	0.933014	0.893581
20	finishedsquarefeet15	1117	0.890750	0.960521
37	pooltypeid7	1066	0.850080	0.815043
33	poolcnt	1052	0.838915	0.801706
54	numberofstories	1032	0.822967	0.772141
27	garagetotalsqft	1017	0.811005	0.668380
26	garagecarcnt	1017	0.811005	0.668380
7	airconditioningtypeid	973	0.775917	0.681185
44	regionidneighborhood	676	0.539075	0.601086

Here, one more question is being raised regarding null values: if Zillow.com has limited resources when they are trying to reduce the influence of the null values, which information and attributes should they try to gather?

Figure 21 correlations of features after dropping all null values



In order to answer this question, I go back to the previous procedure, and change the null value cleaning strategy from fulfill with mean values to drop any null values, by doing this, I got a dataset shape of (50442, 37) and got the following heatmap showing the correlations. However, from Figure 19, I found that after dropping all null values, the log error may seem to have no such strong correlations with other attributes which means I need further study with other methods.

4 Analysis Approach and Solutions

4.1 Time-Series

As the log errors come with the time stamps, therefore, the time-series analysis can be a good way to explore. Start from the correlations between the log errors and the month or days. However, from figure 22, I find that there are no significant relations between the log errors and the month or days.

Figure 22 log errors verses the month or days

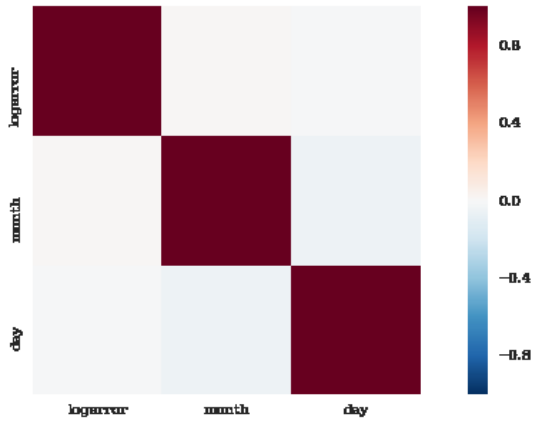
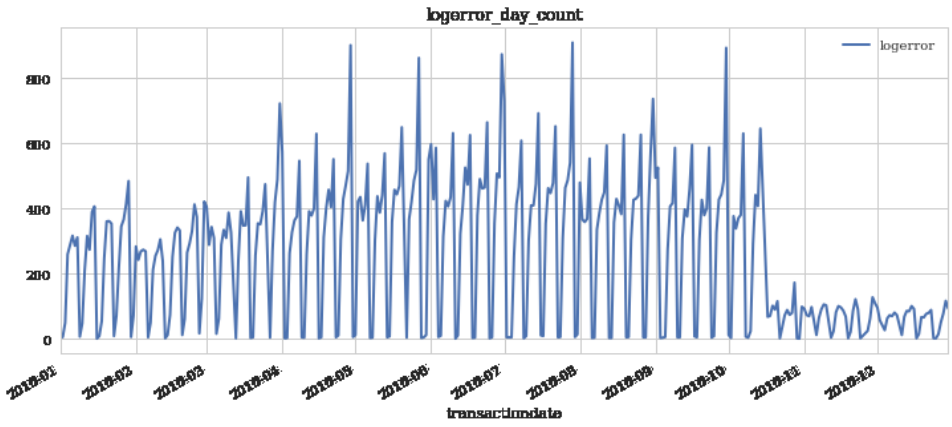


Figure 23 Transaction count versus calendar day



From Figure 23 and Figure 24 which plots the relationships between the log errors and the time, I found that, firstly, the houses more likely to be sold in summer and in winter, especially, during the

thanksgiving and Christmas holiday, which are November and December, the sales amount of houses decreases to a great volumes while for the mean value of log error, the log error faces a very high peak around June, however, during the time, the housing market seems to very active so the lack of transaction records should not be a reason for the high log error. By looking at Figure 25, which is the heatmap plotting the Log Error mean value, it is obvious that in June 11th, there is a very dark mark and this signs a log error which is much higher than mean value and should be analyzed carefully. To conclude, the lack of transactions may not have relations with the log error, instead, the active market may sometimes beyond the prediction model made by Zillow.com, one more assumption is that when Zillow is making the model, it did not count the time-series factor, but this is not very likely. Therefore, the explanation for this unexpectant can located at the unexpected market fluctuations and the current model seems to fail to predict the market fluctuations.

Figure 24 Log Error mean versus calendar day

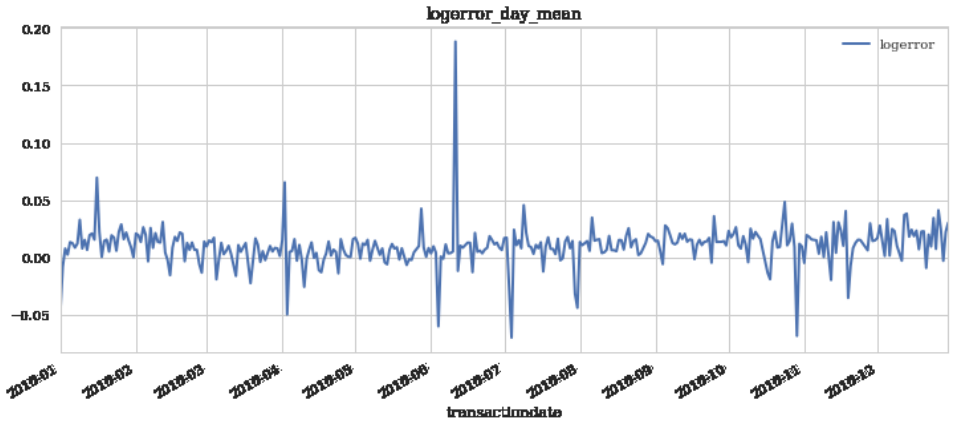
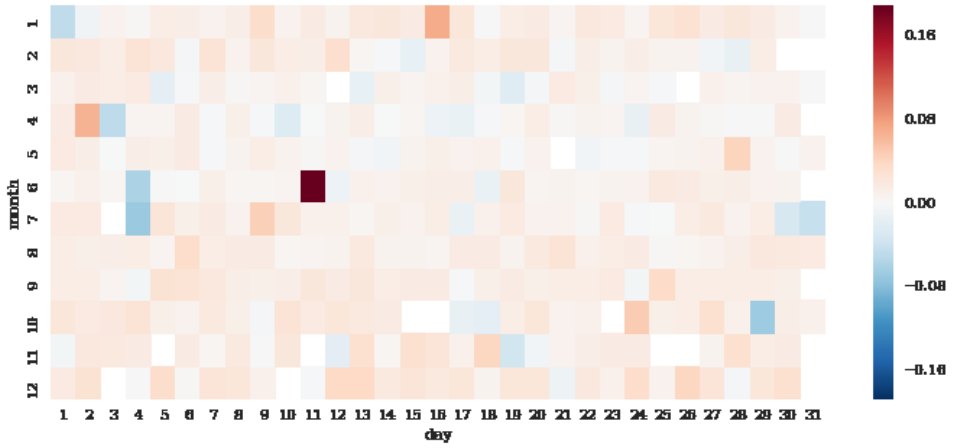
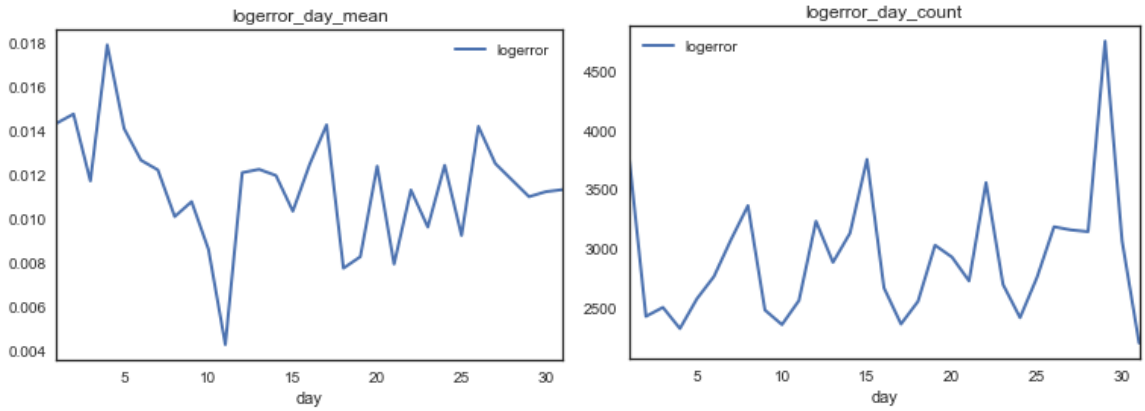


Figure 25 heatmap of Log Error mean



Then, I group the log errors by days. From figure 25 which plot the mean of log errors and count of transactions group by days, from this, I hope to find that is there relationships between the transaction numbers and the log errors, or is there a particular day when the transactions are very likely to happen and one particular day when the log error faces a high or low peak.

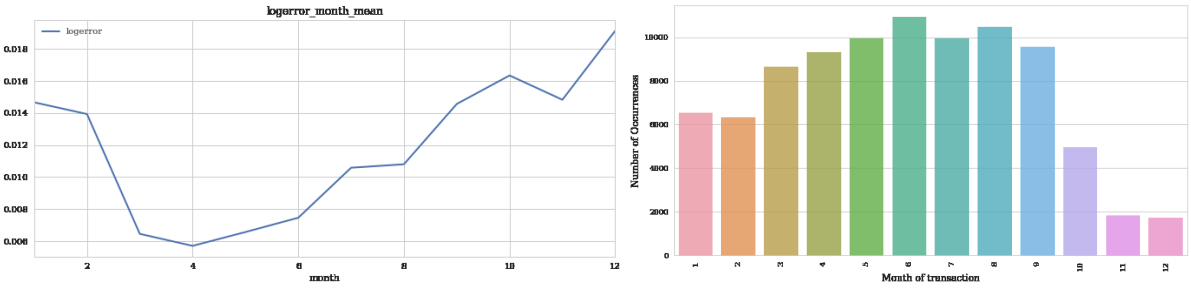
Figure 26 mean of log errors and count of transactions group by days



By reading figure 26, it can be found that the log error does have a high peak around the fifth day for per month and a low peak around the tenth day for each month while houses are most likely to be sold at the end of the month.

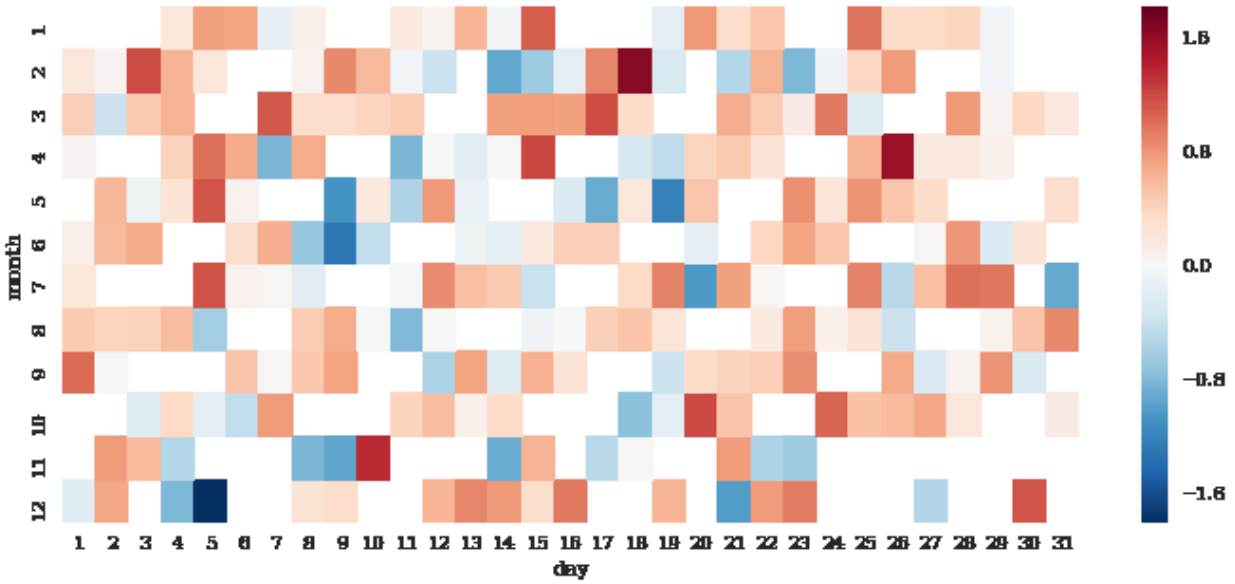
Similarly, for the mean of log errors and count of transactions group by month, it can be assumed that April has the lowest log error and December has the highest log error, this is interesting because from Figure 27, it seems that the log error may has relationship with the number of transactions because in December and November, the number of transactions are the lowest and the log errors are the highest.

Figure 27 mean of log errors and count of transactions group by months



As mentioned above, I am interested in the most inaccurate records and see if there is some pattern or reason saying why these records results in the most inaccurate log errors. Figure 28 shows the heatmap of mean of log errors higher than 0.5 and plot it in the calendar days format. However, this heatmap is relatively random distributed which indicates that the most inaccurate log errors may have no relations with the time.

Figure 28 mean of log errors higher than 0.5



4.2 Linear Regression

Linear regression is a statistical tool used to model the relation between a set of housing characteristics and real estate prices. It estimates the mean value of the response variable, given levels of the predictor variables. The regression approach complements the least squares by identifying how differently real estate prices respond to a change in one unit of housing characteristic, rather than estimating the constant regression coefficient representing the change in the response variable produced by a one-unit change in the predictor variable associated with that coefficient. It estimates the implicit price for each characteristic across the distribution of prices and allows buyers of higher-priced properties to behave differently from buyers of lower priced properties, even if they are within one single housing estate. Thus, it provides a better explanation of the real-world phenomenon and offers a more comprehensive picture of the relationship between housing characteristics and prices.

By selecting all features, the linear regression model's Mean Squared Error is 0.0252320042474, R-Squared is 0.00762250148521 and training Score is 0.00762250148521which is not satisfying. Though, surprisingly, the test Mean Squared Error is 0.0263626375513 and Testing Score is 1.0, this model is not recommended.

Meanwhile, this model takes too much calculation capacity to calculate, therefore, trying to selecte seven feature including Varibles : 'structuretaxvaluedollarcnt', 'bedroomcnt', 'bathroomcnt','fullbathcnt', 'calculatedbathnbr', ' calculatedfinishedsquarefeet', 'finishedsquarefeet12': from the above Correlation Analysis, I assume that these features are the most important and should be concluded when I am trying to reduce the variables.

```
Intercept: -0.0113448510998
structuretaxvaluedollarcnt: -1.34391374427e-08
bedroomcnt: 0.000145144139424
bathroomcnt: 0.0014468285103
fullbathcnt: 0.00171114719759
calculatedbathnbr: -0.00466683863235
calculatedfinishedsquarefeet: 3.28976080942e-06
finishedsquarefeet12: 7.88242488792e-06
Mean Squared Error: 0.0253764551179
R-Squared: 0.00194123288302
Training Score: 0.00194123288302
```

It seems that the scores and the MSEs are not decreasing significantly and this indicates that reducing variables is important and realizable.

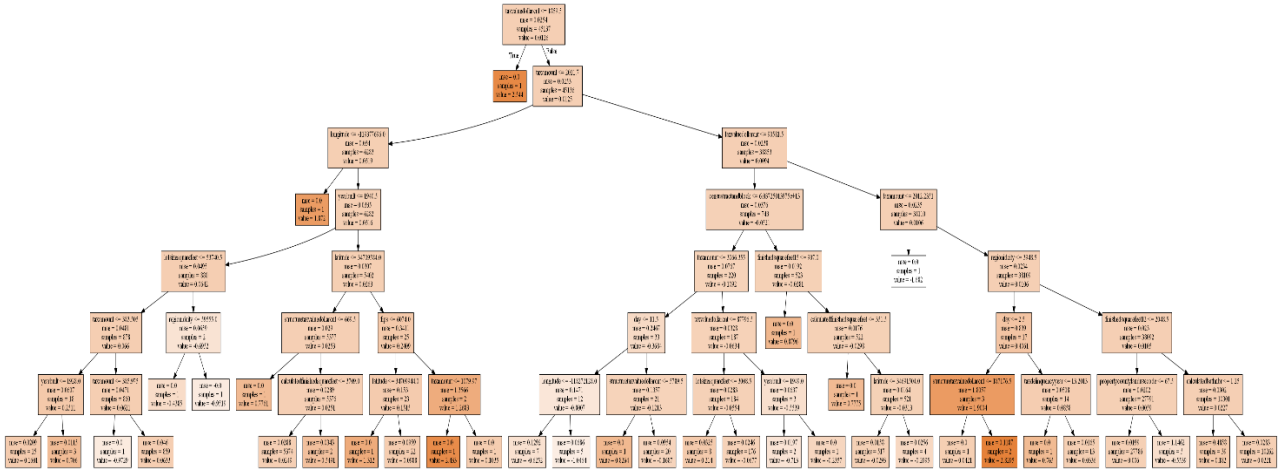
4.3 Regression Trees

The second model is regression trees, firstly, selecting all features,

```
DecisionTreeRegressor(criterion='mse', max_depth=7, max_features=None,
           max_leaf_nodes=None, min_impurity_split=1e-07,
           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, presort=False, random_state=23,
           splitter='best')
```

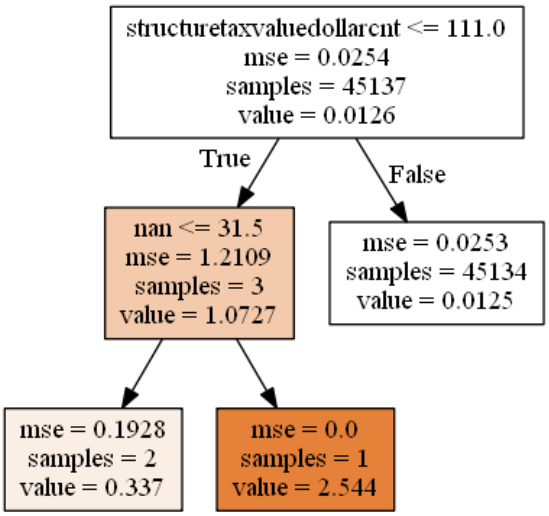
It can be found that the mean squared error is 0.027365346830656577 which is slightly higher than linear regression and Figure 29 shows the respective regression tree.

Figure 29



By selecting eight feature including Varibles :'structuretaxvaluedollarcnt', 'bedroomcnt', 'bathroomcnt','fullbathcnt', 'calculatedbathnbr', ' calculatedfinishedsquarefeet', 'finishedsquarefeet12', nan, the mean squared error is still acceptable: 0.026475205920778618 and figure 30 shows the respective regression tree.

Figure 30



4.4 Reducing Variances

4.4.1 Bagging and Random Forest

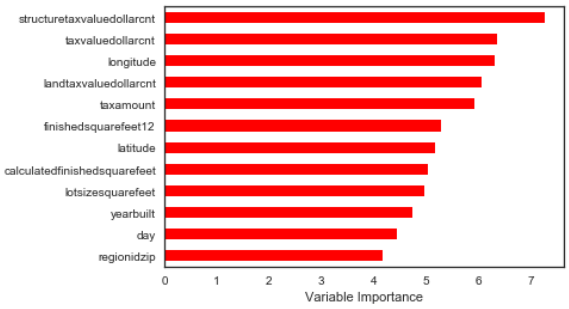
Bootstrap aggregation, or bagging, is a general-purpose procedure for reducing the variance of a statistical learning method. Averaging a set of observations reduces variance. Hence a natural way to reduce the variance and hence increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions. In the classification situation, there are a few possible approaches. For a given test observation, I am going to record the class predicted by each of the B trees, and take a majority vote: the overall prediction is the most commonly occurring majority vote class among the B predictions.

Firstly, I am going to apply the random forest regressor method,

```
RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=None,
           max_features=19, max_leaf_nodes=None, min_impurity_split=1e-07,
           min_samples_leaf=1, min_samples_split=2,
           min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
           oob_score=False, random_state=23, verbose=0, warm_start=False)
```

Firstly, trying to set Max features as 19 and the following mean squared error is 0.0293293268240374. Then setting Max features as 15 then mean squared error is 0.0291799843881519.

Figure 31



Features	Importanc	Features	Importanc
structuretaxvaluedollarcnt	7.259378	month	3.422334
taxvaluedollarcnt	6.342753	propertyzoningdesc	3.345104
longitude	6.311709	rawcensustractandblock	3.231073
landtaxvaluedollarcnt	6.046392	regionidneighborhood	2.410675
taxamount	5.917645	regionidcity	2.326605
finishedsquarefeet12	5.281085	nan	2.228888
latitude	5.160203	censustractandblock	1.967576
calculatedfinishedsquarefe	5.041697	bedroomcnt	1.648505
lotsizesquarefeet	4.961062	propertycountylandusecoc	1.508768
yearbuilt	4.747469	bathroomcnt	1.49255
day	4.430885	calculatedbathnbr	1.172267
regionidzip	4.168153		

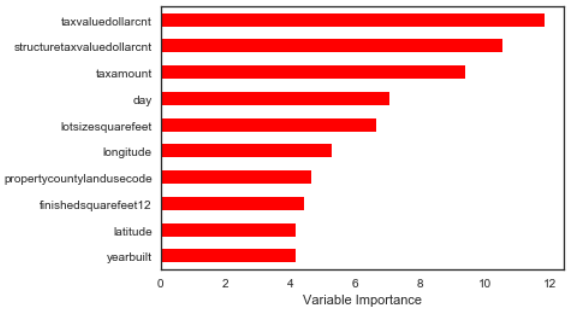
Figure 31 shows the importance of each feature, and the structure tax value amount and tax value dollar amount are the two most important features.

4.4.2 Boosting

Key difference between boosting and random forests: in boosting, the growth of a particular tree takes into account the other trees that have already been grown. This often results in using smaller trees, which aids interpretability of the model.

```
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
             learning_rate=0.01, loss='ls', max_depth=3, max_features=None,
             max_leaf_nodes=None, min_impurity_split=1e-07,
             min_samples_leaf=1, min_samples_split=2,
             min_weight_fraction_leaf=0.0, n_estimators=500,
             presort='auto', random_state=23, subsample=1.0, verbose=0,
             warm_start=False)
```

Figure 32



Features	Importanc	Features	Importanc
taxvaluedollarcnt	11.819306	regionidcity	3.207684
structuretaxvaluedollarcnt	10.532181	regionidzip	3.066743
taxamount	9.395251	finishedsquarefeet6	2.872364
day	7.03664	regionidneighborhood	1.865949
lotsizesquarefeet	6.655465	finishedsquarefeet15	1.667827
longitude	5.264994	calculatedbathnbr	1.508694
propertycountylandusecod	4.631957	rawcensustractandbloc	1.418038
finishedsquarefeet12	4.421881	nan	1.322695
latitude	4.174443	bathroomcnt	1.264389
yearbuilt	4.146329	month	1.171287
calculatedfinishedsquarefe	3.731673		

The Mean squared error for boosting is 0.026359714940661735

Figure 32 shows the importance of each feature, and the structure tax value amount and tax value dollar amount are the two most important features.

4.4.3 Subset Selection

This involves identifying a subset of available predictors that are related to the response variable.

To identify a best model, I defined a metric such as the smallest RSS, and pick among the models using prediction error (MSE), or other metrics such as adjusted R-squared, AIC, or BIC.

Instead of searching all predictors at once, I utilize stepwise selection, which is computationally more efficient and yields a smaller set of useful predictors.

Figure 33

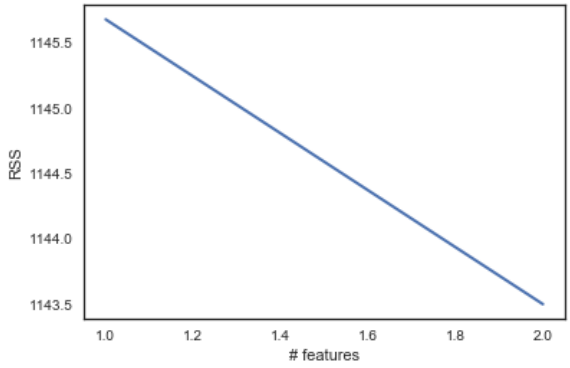
OLS Regression Results

Dep. Variable:	logerror	R-squared:	0.010
Model:	OLS	Adj. R-squared:	0.010
Method:	Least Squares	F-statistic:	222.8
Date:	Wed, 06 Dec 2017	Prob (F-statistic):	4.93e-97
Time:	16:48:07	Log-Likelihood:	18906.
No. Observations:	45137	AIC:	-3.781e+04
Df Residuals:	45135	BIC:	-3.779e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
finishedsquarefeet12	1.176e-05	6.18e-07	19.037	0.000	1.05e-05	1.3e-05
taxamount	-1.22e-06	1.32e-07	-9.270	0.000	-1.48e-06	-9.62e-07

Omnibus:	40611.110	Durbin-Watson:	1.983
Prob(Omnibus):	0.000	Jarque-Bera (JB):	37379669.074
Skew:	3.238	Prob(JB):	0.00
Kurtosis:	143.831	Cond. No.	7.78

Figure 34



From Figure 33 and Figure 34, it can be observed that finished square feet 12 and tax amount are the two most important variables Subset Selection choose.

4.4.4 Validation Set and Cross-Validation

From the validation perspective, it can be founded that the best 2-predictor model using best subset selection & validation approach choose finished square feet 12 and tax amount as the two most important variables.

```
Best 2-predictor model using best subset selection & validation approach:
finishedsquarefeet12    1.140823e-05
taxamount              -9.465980e-07
dtype: float64
```

Figure 35Forward Validation

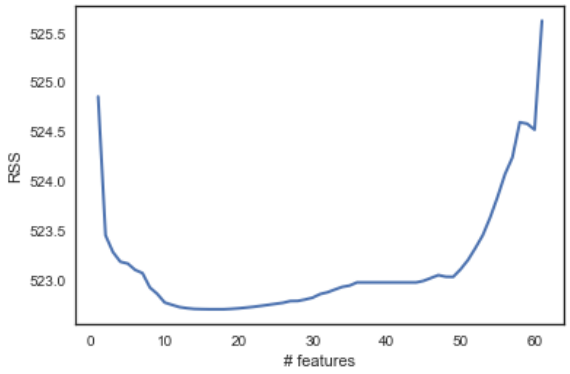


Figure 36Backward Validation

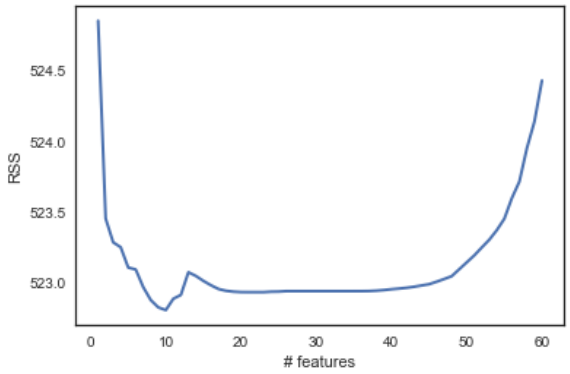
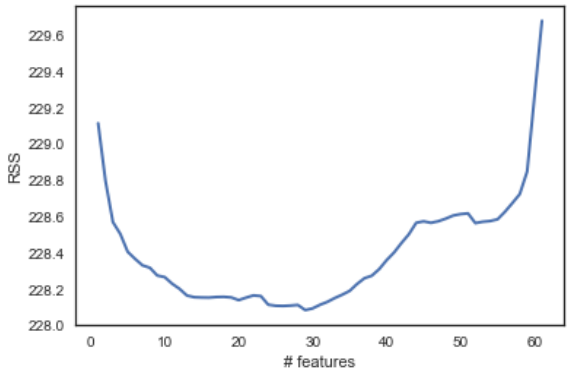


Figure 37K-fold Validation



OLS Regression Results

Dep. Variable:	logerror	R-squared:	0.011
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	41.78
Date:	Wed, 06 Dec 2017	Prob (F-statistic):	4.74e-83
Time:	17:05:27	Log-Likelihood:	14906.
No. Observations:	36091	AIC:	-2.979e+04
Df Residuals:	36081	BIC:	-2.971e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
finishedsquarefeet12	8.603e-06	3.61e-06	2.385	0.017	1.53e-06	1.57e-05
taxamount	-3.563e-06	3.89e-07	-9.155	0.000	-4.33e-06	-2.8e-06
taxvaluedollarcnt	3.139e-08	4.79e-09	6.560	0.000	2.2e-08	4.08e-08
garagetotalsqft	-1.364e-05	7.17e-06	-1.904	0.057	-2.77e-05	4.04e-07
regionidcounty	2.753e-06	9.6e-07	2.868	0.004	8.72e-07	4.64e-06
garagecarcnt	-0.0015	0.002	-0.709	0.478	-0.006	0.003
regionidcity	-2.226e-08	1.78e-08	-1.248	0.212	-5.72e-08	1.27e-08
fullbathcnt	-0.0008	0.001	-0.592	0.554	-0.004	0.002
calculatedfinishedsquarefeet	5.276e-06	3.66e-06	1.442	0.149	-1.89e-06	1.24e-05
propertycountylandusecode	-0.0003	0.000	-1.661	0.097	-0.001	4.73e-05

Omnibus:	31639.513	Durbin-Watson:	1.984
Prob(Omnibus):	0.000	Jarque-Bera (JB):	28372201.773
Skew:	3.085	Prob(JB):	0.00
Kurtosis:	140.219	Cond. No.	1.92e+06

By observing the RSS (residual sum of squares), when choosing around 10 features, the models have the lowest RSS, meanwhile, the finished square feet 12, tax amount, tax value amount, garage total square feet should be considered as important features and when choosing around 10 features, the models have the lowest sum of bias and variance which prevents the overfitting.

5 Conclusion

In general, this study is interested in how accurate are estimates of housing value. Reliable estimates of value are needed for a number of reasons. For example, tax assessors, appraisers, and real estate agents require reasonably accurate estimates to perform their jobs. One estimate of value that has received ongoing interest in real estate literature is the estimate of value provided by homeowners. Homeowners have been asked to determine the value of their homes in previous studies and in the American Housing Survey conducted by the U.S.

The results show that Zillow's Z-estimate is satisfying and though mean value of the log error is 0.0114 and is higher than zero, 50% of log errors are in the range of -0.0253 to 0.0392 which indicates that if a house is worth $1,000,000 then averagely, the Z-estimates will be 1026597.02175 and the range of the Z-estimate is 50% confident in the range of 943408.967323 to 1094460.26782. Real estate professionals sometimes get inquiries from prospective real estate buyers and sellers about the Zestimate. Understanding how the Zestimate is calculated, along with its strengths and weaknesses, can provide potential customers with an opportunity to make a better decision.

The other findings are: the Zestimate's accuracy depends on location and availability of data in an area. Some houses have deeply detailed information on homes such as number of bedrooms, bathrooms and square footage and others do not. Under 30 available features, the more data available, the more accurate the Zestimate value, with more than 30 available features, this argument may be flawed.

To improve Zestimate accuracy, Zillow allow homeowners to edit their home facts and then incorporate this information into Zestimate calculations but if Zillow has limited resources, then finished square feet 12, tax amount, tax value amount, garage total square feet should be gathered and discovered firstly.

Though this project uses the housing price data of Southern California only and may not present the entire country, but this could be a start to analyze the housing market and with more data, better algorithms, better calculation capability, the model can be adjusted better and produce a more precise prediction to predict the housing market for the entire country.

References

Case, K., & Shiller, R. (1988). The Efficiency of the Market for Single-Family Homes. doi:10.3386/w2506

Chen, Z., Cho, S., Poudyal, N., & Roberts, R. K. (2009). Forecasting Housing Prices under Different Market Segmentation Assumptions. *Urban Studies*, 46(1), 167-187. doi:10.1177/0042098008098641

Corcoran, C., & Liu, F. (2014). Accuracy of Zillow's Home Value Estimates. *Real Estate Issues*, 39, 45-49.

Epley, D. (2016). Assumptions And Restrictions On The Use Of Repeat Sales To Estimate Residential Price Appreciation. *Journal of Real Estate Literature*, 24, 275-284.

Green, R. K., Malpezzi, S., & Mayo, S. K. (2005). Metropolitan-Specific Estimates of the Price Elasticity of Supply of Housing, and Their Sources. *American Economic Review*, 95(2), 334-339. doi:10.1257/000282805774670077

Groen, A. (n.d.). Simple Starter - RandomForest Regressor. Retrieved November 30, 2017, from https://www.kaggle.com/arjanso/simple-starter-randomforest-regressor

Guerrieri, V., Hartley, D., & Hurst, E. (2010). Endogenous Gentrification and Housing Price Dynamics. doi:10.3386/w16237

H. (n.d.). A chinese translation of material. Retrieved November 30, 2017, from https://www.kaggle.com/qwert1234/a-chinese-translation-of-material-updating

Hollas, D. R., Rutherford, R. C., & Thomson, T. A. (2010). Zillow's estimates of single-Fmaily Housing Values. The Appraisal Journ*al*, 26-32.

Mak, S., Choy, L., & Ho, W. (2010). Quantile Regression Estimates of Hong Kong Real Estate Prices. *Urban Studies*, 47(11), 2461-2472. doi:10.1177/0042098009359032

Montero, J. M., & Larraz, B. (2009). Estimating Housing Prices: A Proposal with Spatially Correlated Data. *International Advances in Economic Research*, 16(1), 39-51. doi:10.1007/s11294-009-9244-5

Mueller, J. M., & Loomis, J. B. (2014). Does the estimated impact of wildfires vary with the housing price distribution? A quantile regression approach. *Land Use Policy*, 41, 121-127. doi:10.1016/j.landusepol.2014.05.008

Nevia, S. (n.d.). Zillow's Home Value Exploratory Data Analysis. Retrieved November 30, 2017, from https://www.kaggle.com/neviadomski/zillow-s-home-value-exploratory-data-analysis

Nong, K. S. (n.d.). Simple EDA Geo Data & Time Series. Retrieved November 30, 2017, from https://www.kaggle.com/kueipo/simple-eda-geo-data-time-series/notebook

R., T., & G. (n.d.). Log errors of nearest neighbors and some eda. Retrieved November 30, 2017, from https://www.kaggle.com/rteja1113/log-errors-of-nearest-neighbors-and-some-eda

Sah, V., Conroy, S. J., & Narwold, A. (2015). Estimating School Proximity Effects on Housing Prices: the Importance of Robust Spatial Controls in Hedonic Estimations. *The Journal of Real Estate Finance and Economics*, 53(1), 50-76. doi:10.1007/s11146-015-9520-5

Srinivasan, V. (n.d.). Zillow EDA On Missing Values & Multicollinearity. Retrieved November 30, 2017, from https://www.kaggle.com/viveksrinivasan/zillow-eda-on-missing-values-multicollinearity

Wang, Y., Wang, S., Li, G., Zhang, H., Jin, L., Su, Y., & Wu, K. (2017). Identifying the determinants of housing prices in China using spatial regression and the geographical detector technique. *Applied Geography*, 79, 26-36. doi:10.1016/j.apgeog.2016.12.003

Wu, T., Cheng, M., & Wong, K. (2017). Bayesian analysis of Hong Kong's housing price dynamics. *Pacific Economic Review*, 22(3), 312-331. doi:10.1111/1468-0106.12232

Xu, Y. (2011). Using Repeat Sales Model to Estimate Housing Index in Price Engineering. *Systems Engineering Procedia*, 2, 33-39. doi:10.1016/j.sepro.2011.10.005

Yang, Y., Liu, J., Xu, S., & Zhao, Y. (2016). An Extended Semi-Supervised Regression Approach with Co-Training and Geographical Weighted Regression: A Case Study of Housing Prices in Beijing. *ISPRS International Journal of Geo-Information*, 5(1), 4. doi:10.3390/ijgi5010004.