DSC 520 Final Project

Name: Edris Safari
Date: 02/25/2020
Title: Real Estate Data Analysis

Section 1 – Getting Started

In this section, we will describe three data sets that will be examined for this project. Datasets are related in the Real Estate industry but vary in content. Each dataset will be examined according to their content. A combined report will detail the findings of each dataset.

Dataset 1

Data Source

https://www.kaggle.com/samdeeplearning/vt-nh-real-estate

Description

This dataset contains features of houses in three towns in Vermont, which make up a sizable chunk of the real estate firm's business. The dataset is divided into test, train and validate data sets with test having 24 rows, train 138 and validate with 70 rows. There are 28 column describing features such as number of bedrooms, yard size, etc.

Goal

We will try to cross validate the results between Train, Validate, and, Test. We will select appropriate independent variables after exploratory data analysis and run a regressions model which we will test against the test dataset and then validate.

Dataset 2

Data Source

https://www.kaggle.com/quantbruce/real-estate-price-prediction

Description

Dataset columns are 'transaction_date' , 'house_age' , 'distance_to_the_nearest_MRT_station' , 'number_of_convenience_stores' , 'latitude' , 'longitude' , 'house_price_of_unit_area'. There are 500 records in the dataset.

Goal

Evaluate Correlation between independent variables and make prediction on the dependent variable 'house_price_of_unit_area'

Dataset 3

Data Source

https://www.kaggle.com/c/zillow-prize-1

https://www.kaggle.com/philippsp/exploratory-analysis-zillow

Description

There are two data sets with over 1 million records each and 58 columns. properties_2016 and properties_2017 datasets contain data for each year.

Goal

Reduce error between actual home price and zillow's estimate. We will perform EDA and correlation analysis and create a model to examine variation in estimates vs. actual home price.

Needed packages

Packages needed for calculation, analysis and plotting the data sets are listed below:

```
library(data.table)
library(dplyr)
library(ggplot2)
library(stringr)
library(DT)
library(tidyr)
library(corrplot)
library(leaflet)
library(lubridate)
```

Plots and tables

We will create histograms and density plots, scatter plots and correlation plots to examine and understand the data.

Questions for hypothesis testing

1. Does location determine price?
2. Which affect the price of a home? Number of rooms and the square foot?
3. How Does age affect the price of a home?
4. What are the common factors among certain price range?

Section 2 –Cleaning Data and Exploratory Data Analysis

In this section, we select a dataset from the previous section. We will then import the dataset, perform preliminary evaluation of the dataset and prepare the dataset for the next phase which will be to test a few hypotheses, perform correlation analysis and select features to perform a regression model.

Importing and cleaning the data sets:

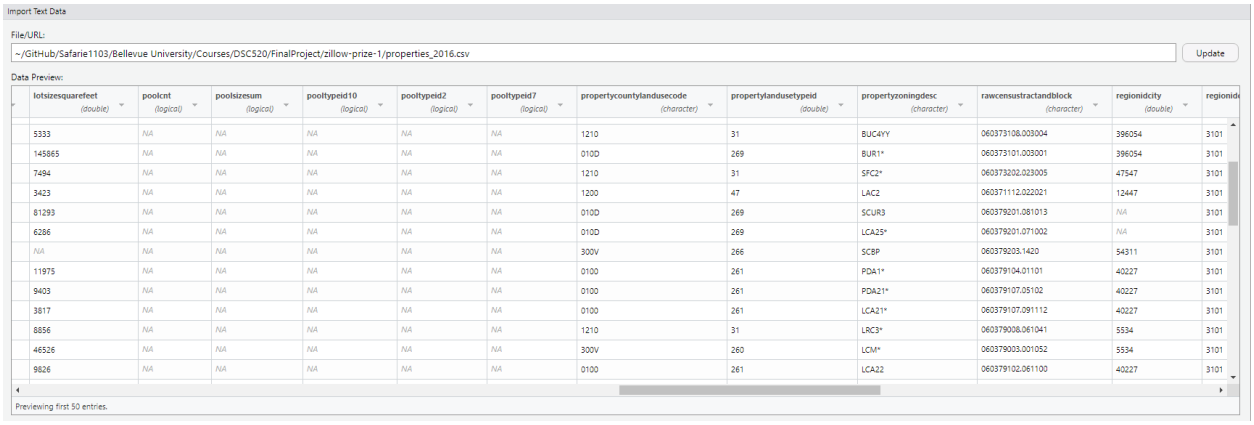
There are two data sets we will import for analysis:

- properties <- read.csv("zillow-prize-1/properties_2016.csv")

  This data set in indexed by attribute **parceled** and the following attributes:

```
> names(properties)
 [1] "parcelid"                  "airconditioningtypeid"     "architecturalstyletypeid"  "basementsqft"
 [5] "bathroomcnt"               "bedroomcnt"                "buildingclasstypeid"       "buildingqualitytypeid"
 [9] "calculatedbathnbr"         "decktypeid"                "finishedfloor1squarefeet"  "calculatedfinishedsquarefeet"
[13] "finishedsquarefeet12"      "finishedsquarefeet13"      "finishedsquarefeet15"      "finishedsquarefeet50"
[17] "finishedsquarefeet6"       "fips"                      "fireplacecnt"              "fullbathcnt"
[21] "garagecarcnt"              "garagetotalsqft"           "hashottuborspa"            "heatingorsystemtypeid"
[25] "latitude"                  "longitude"                 "lotsizesquarefeet"         "poolcnt"
[29] "poolsizesum"               "pooltypeid10"              "pooltypeid2"               "pooltypeid7"
[33] "propertycountylandusecode" "propertylandusetypeid"     "propertyzoningdesc"        "rawcensustractandblock"
[37] "regionidcity"              "regionidcounty"            "regionidneighborhood"      "regionidzip"
[41] "roomcnt"                   "storytypeid"               "threequarterbathnbr"       "typeconstructiontypeid"
[45] "unitcnt"                   "yardbuildingsqft17"        "yardbuildingsqft26"        "yearbuilt"
[49] "numberofstories"           "fireplaceflag"             "structuretaxvaluedollarcnt" "taxvaluedollarcnt"
[53] "assessmentyear"            "landtaxvaluedollarcnt"     "taxamount"                 "taxdelinquencyflag"
[57] "taxdelinquencyyear"        "censustractandblock"
```

A sample of data is shown below:



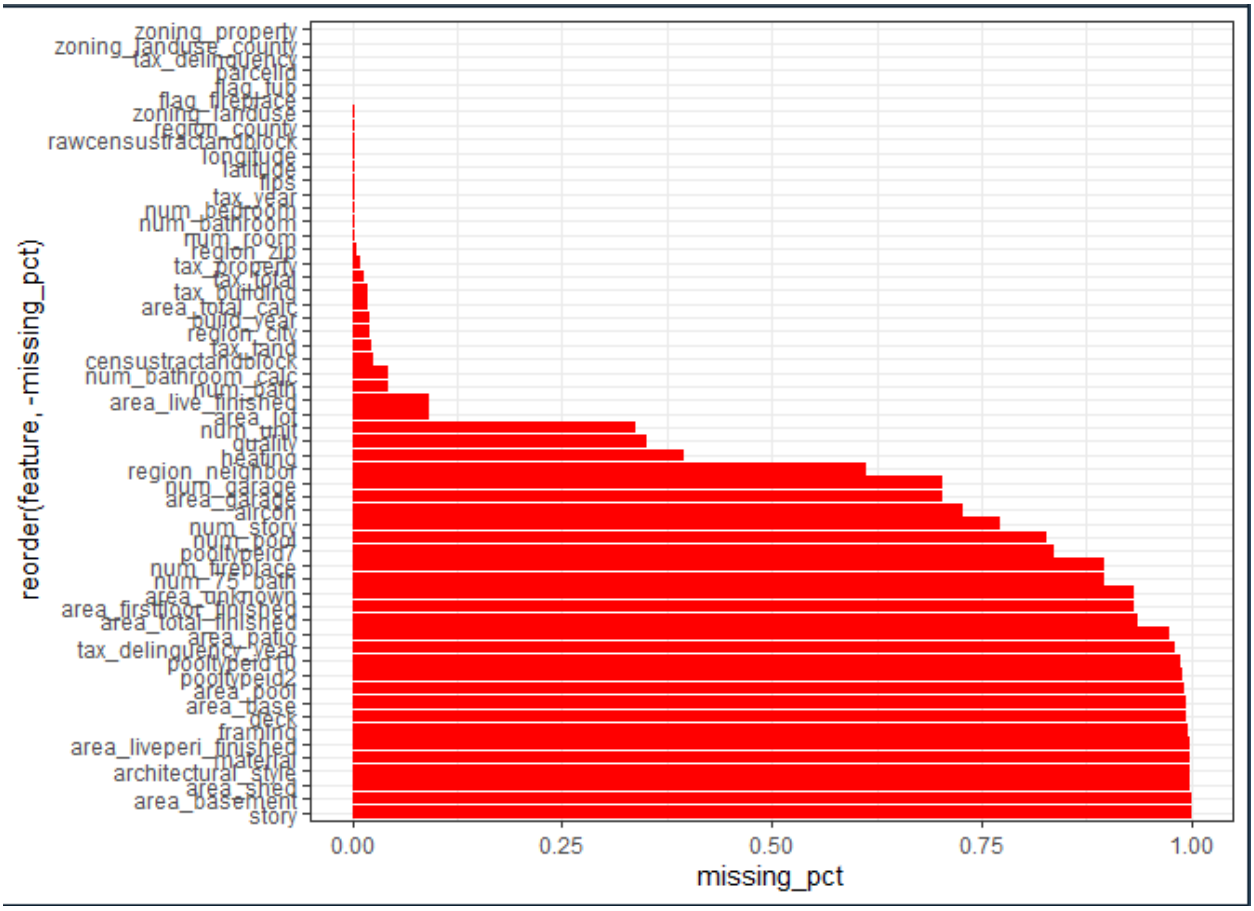
As shown, there are columns with numerous NA's. The following code identifies columns with NA's and allows us to select columns with less than 75% NA's.

```
missing_values <- properties %>% summarize_each(funs(sum(is.na(.))/n()))
missing_values
missing_values <- gather(missing_values, key="feature", value="missing_pct")
missing_values %>%
  ggplot(aes(x=reorder(feature,-missing_pct),y=missing_pct)) +
  geom_bar(stat="identity",fill="red")+
  coord_flip()+theme_bw()

good_features <- filter(missing_values, missing_pct<0.75)

good_features
```

The ggplot graph below shows the percentage of NA's in each column.



We create a data frame that contains less than 75% NA's. The **good_features** dataframe has two columns-'features' and 'missing_pct'. The 'features' column contains the list of columns with less than 75% NA's. We would still have to handle the NA's which may appear in rows of the dataset, but this code guarantees that we remove columns with more that 75% NA's because they would adversely affect our analysis.

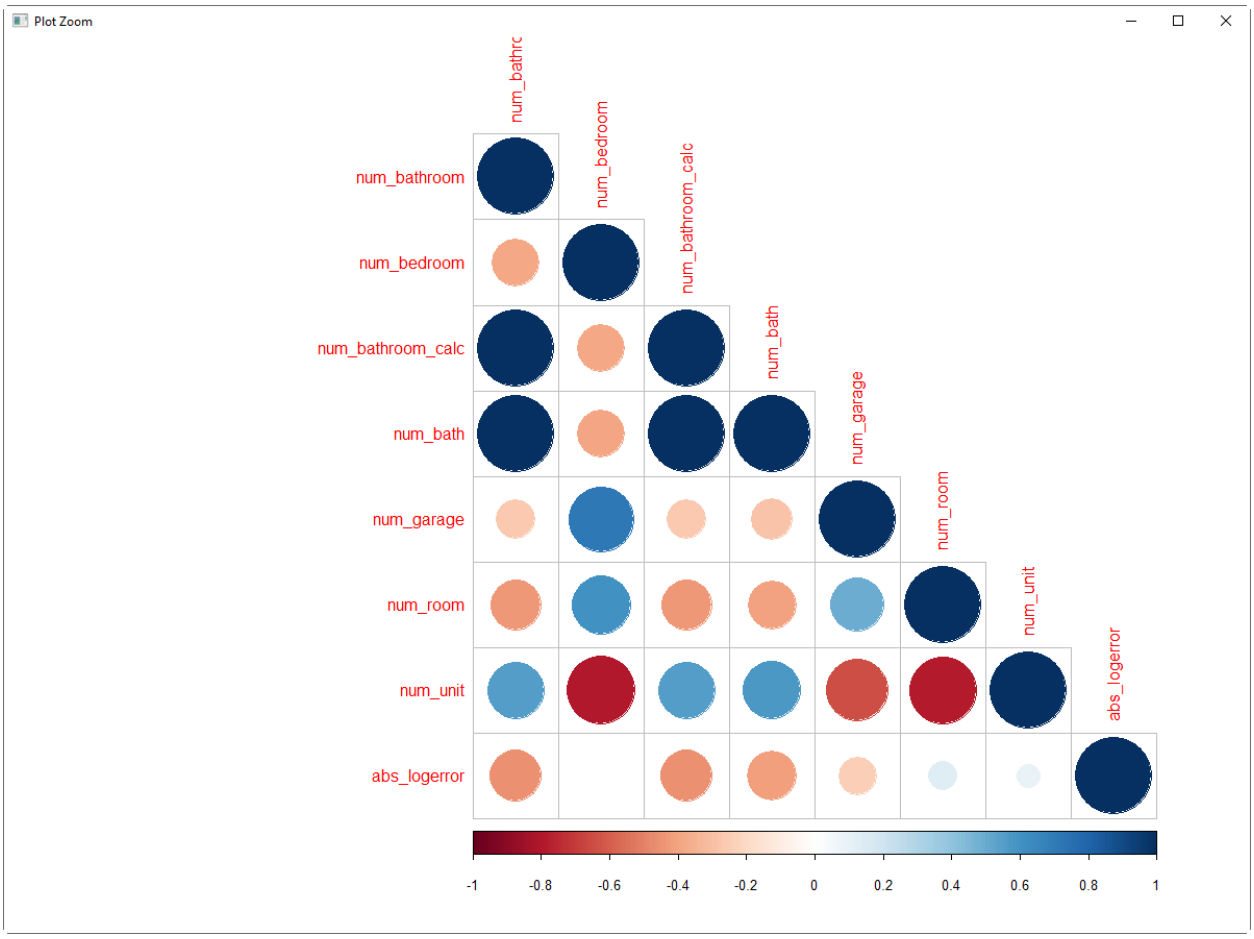
We will use the features in **good_features** dataframe for our analysis. For example, the code below identifies the feature with '_num' in them:

```
vars <- good_features$feature[str_detect(good_features$feature,'num_')]
```

, and code below creates the correlation coefficient diagram.

```
cor_tmp <- transactions %>% left_join(properties, by="parcelid")
tmp <- cor_tmp %>% select(one_of(c(vars,"abs_logerror")))

corrplot(cor(tmp, use="complete.obs"),type="lower")
```



- transactions <- read.csv("zillow-prize-1/train_2016.csv")

  This dataset is also indexed by **parceled** and has additional features listed below:

```
> names(transactions)
[1] "parcelid"        "logerror"        "transactiondate"
> |
```

  A sample of data is shown below:

Import Text Data

File/URL:

~/GitHub/Safarie1103/Bellevue University/Courses/DSC520/FinalProject/zillow-prize-1/train_2016.csv

Data Preview:

parcelid (double)	logerror (double)	transactiondate (double)
11016594	0.0276	2016-01-01
14366692	-0.1684	2016-01-01
12098116	-0.0040	2016-01-01
12643413	0.0218	2016-01-02
14432541	-0.0050	2016-01-02
11509835	-0.2705	2016-01-02
12286022	0.0440	2016-01-02
17177301	0.1638	2016-01-02
14739064	-0.0030	2016-01-02
14677559	0.0843	2016-01-03
10854446	0.3825	2016-01-03

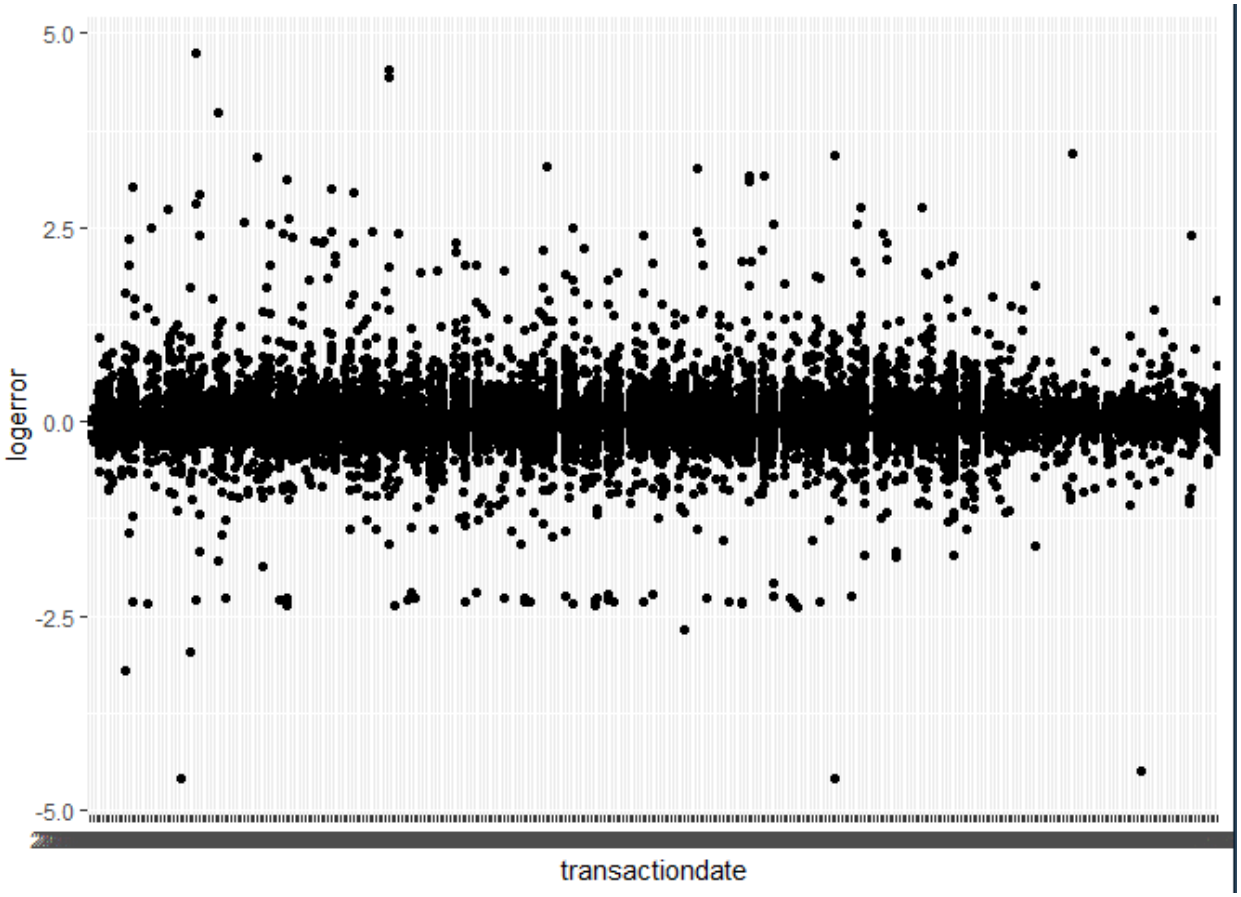
This dataset doesn't need to be cleaned. The **parcelid** of this dataset and **parcelid** of the 'properties' dataset is related and joining them would show the transaction date and the log error which is the error between the sales price of the property and the estimated price. This log error will be part of the analysis as it shows the estimate vs actual sales price. We may want to investigate the variations in the logerror and find correlations.

Section 3 –Writeups

- Discuss how you plan to uncover new information in the data that is not self-evident.

  The transactions dataset records the date of transactions and the log error. The scatter plot below shows the log error distribution. It shows that on some dates the log error is farther away from the mean. We want to plot the logerror in various ways to better understand the variation. Below is the scatter diagram of log error versus transaction date.

```
27  #Scatter plot of log error
28  transactions_2016 %>%
29    ggplot(aes(x=transactiondate,y=logerror)) +
30    geom_point()
```

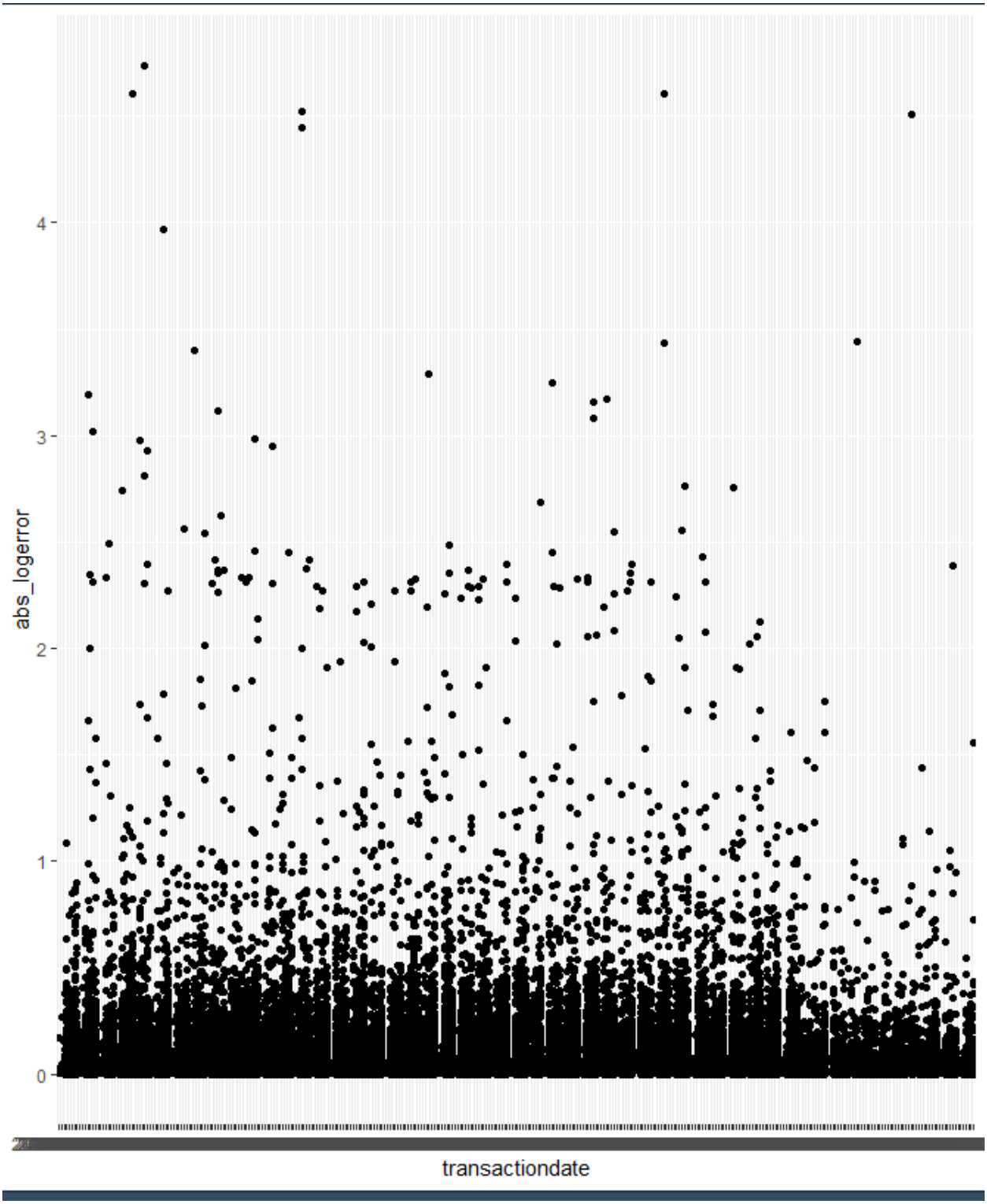


- What are different ways you could look at this data to answer the questions you want to answer?

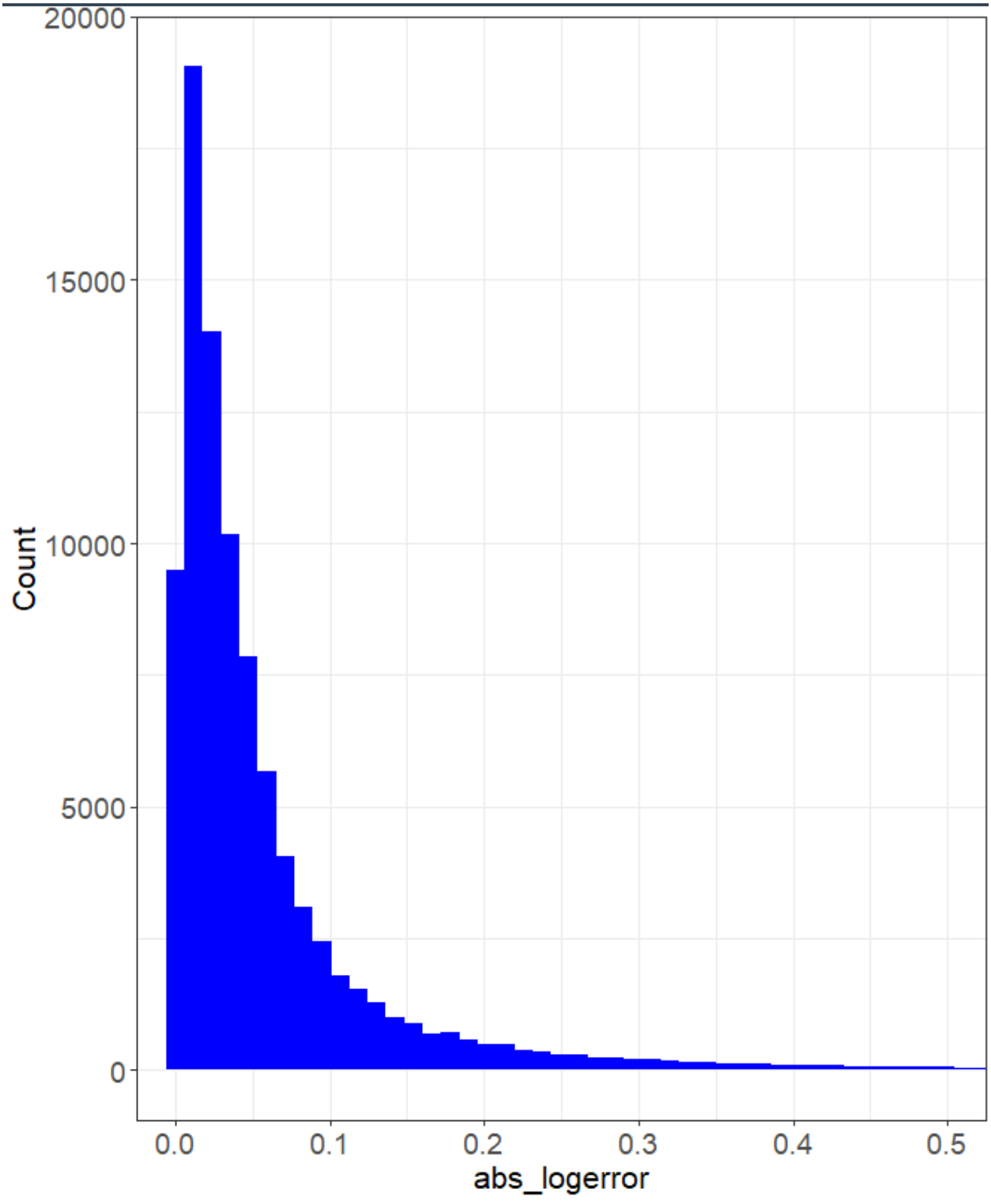
  We want to look at the absolute log error in the transactions dataset. To do this we create a new column in the dataset as shown below:

```
transactions$abs_logerror <- abs(transactions$logerror)
```

  The plot is shown below:



Another way of looking at the absolute log error is with a histogram shown below.



The plot below shows the variation of log error based on transaction month. It looks like the zestimates were a bit off in February and December month. It would be worth seeing if the same behavior exists in the 2017 data set.

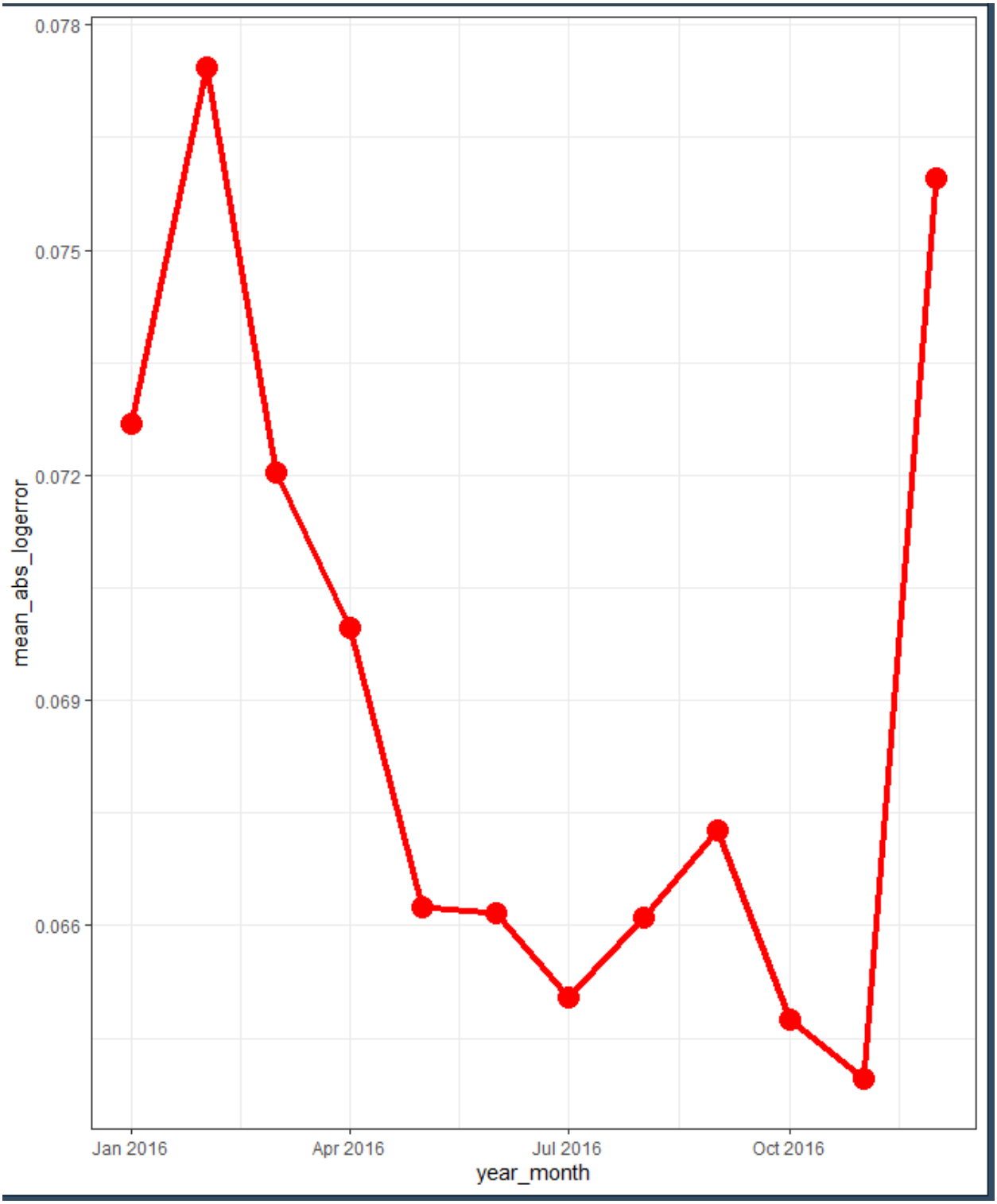
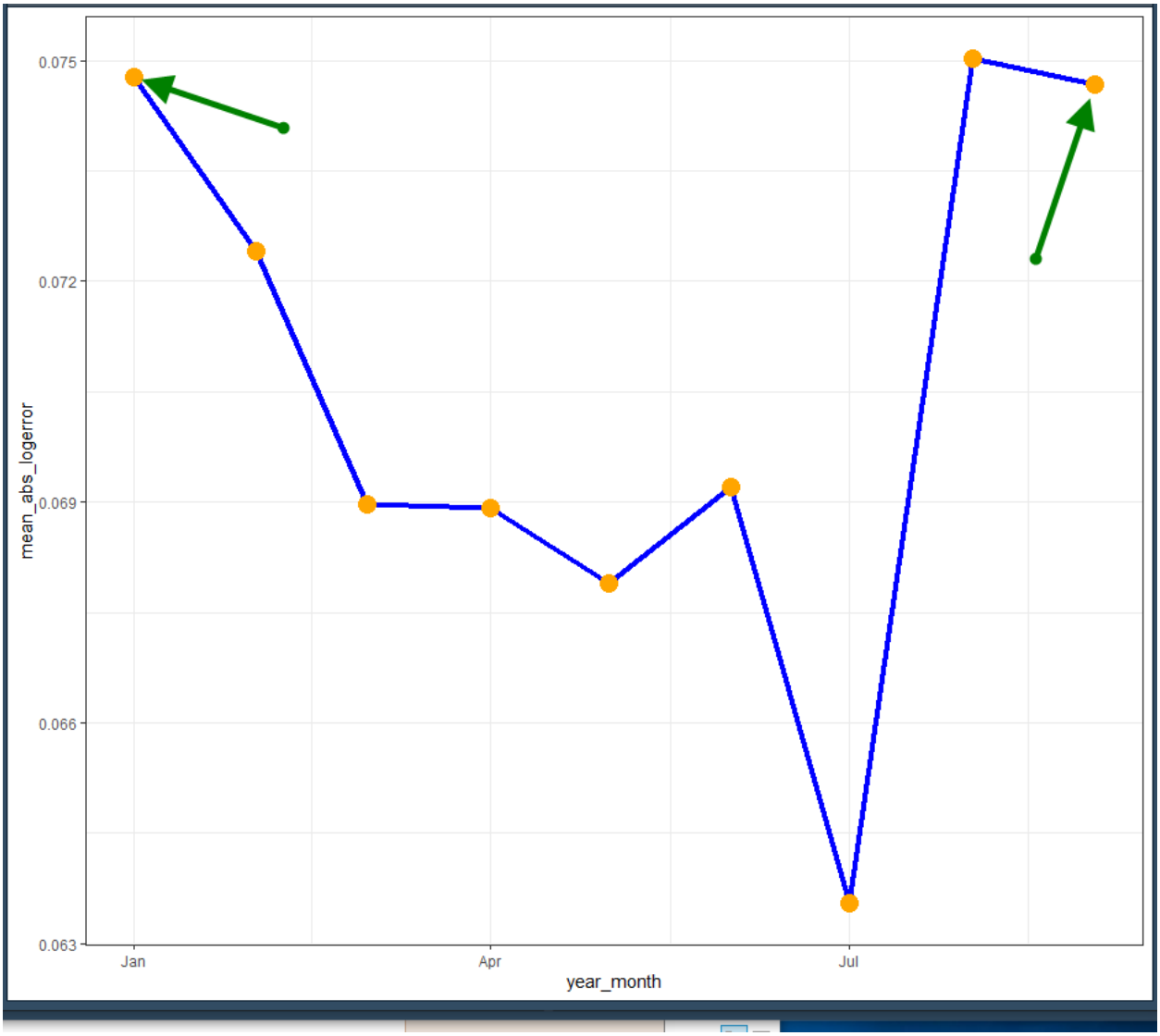


Figure below shows monthly log error for 2017. Notice the spikes in Jan and Dec. and also in November. Log error in Nov 2016 is closer to the mean.



- Do you plan to slice and dice the data in different ways, create new variables, or join separate data frames to create new summary information? Explain.

  The new variables in the transaction dataset were the absolute value of log error and the means absolute value. They helped us create the graphs above and give us insight into the monthly variation of log error.

- How could you summarize your data to answer key questions?

  The key question is what causes the log error to increase. The analysis above only showed monthly variation and did not seem to be correlated. We would merge the transactions and properties dataset and perform correlational analysis for that discovery.

- What types of plots and tables will help you to illustrate the findings to your questions? Ensure that all graph plots have axis titles, legend if necessary, scales are appropriate, appropriate geoms used, etc.).

  Histograms, scatter plots, geom_smooth with linear model.

- What do you not know how to do right now that you need to learn to answer your questions?

  I will have to do regression modeling, so I will resort to lessons in this course.

- Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.

  Time permitting, we intend to fit a regression model and run the model against a test data set. The test dataset will have been created from the existing datasets. We will create a training dataset to create the model and a test dataset to test it. The ration of training vs. test dataset will be 80 to 20 percent.

Section 4 – Final

Introduction

In real estate as in any other commerce that sells a product, the correct price to buy or sell takes center stage. Factors that go into determining the sales prices are more that “location, location, location”. In this study, we examine two datasets. The ‘properties’ data set contains over one million records of properties in Los Angeles, Orange County and Ventura County. The features in this dataset include number of rooms, areas, longitude, latitude and several other features. The ‘transactions’ data set contains less than 100,000 records of sell transactions. The features in this table are date, logerror and pacelid. Parceled can be used to link to the ‘properties’ data set.

Problem statement addressed by this study

The dataset is rich with data about the properties. The transactions dataset shows the logerror which is the error in the estimate. So, the estimate is already made based on the features in the ‘properties’ dataset. The goal of this project is to explore the datasets and show where those estimates were accurate and where they were over or underestimated.

Techniques used to address the problem statement

We added calculated absolute log error to the dataset and plotted it against various features. By creating a percentile category on the absolute log error, we could create charts that showed how each category performed against certain features such as number of bedrooms and total living area.

Analysis

Several factors were found to contribute to the accuracy of the predictions. The underlying algorithm performs good predictions, and to bring the logerror to below the current level will require further analysis, market research, etc.

Implications

The implication of this sort of studies span to areas peripheral to real estate. Lending rates as well as property insurance and other related industries could leverage from these studies.

Limitations

This study does not guarantee sale of the properties at a certain time, but with proper algorithm and supporting data we can calculate probability of such event.

Concluding Remarks

As data is available on many aspects of real estate, the analysis of them can reveal opportunities that may not otherwise be visible. With techniques available to us we can dig deep into data, dissect them, create new data based on existing data, and find patterns in their distribution. From there we can make conclusions