

DSC 520 Final Project Template

Name: Edris Safari

Date: 02/06/2020

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/samdeeeplearning/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>

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 is indexed by attribute **parceled** and the following attributes:

```
> names(properties)
[1] "parcelid" "airconditioningtypeid" "architecturalstyletypeid" "basementsqft"
[5] "bathroomcnt" "bedroomcnt" "buildingclasstypid" "buildingqualitytypeid"
[9] "calculatedbathnbr" "decktypeid" "finishedfloor1squarefeet" "calculatedfinishedsquarefeet"
[13] "finishedsquarefeet12" "finishedsquarefeet13" "finishedsquarefeet15" "finishedsquarefeet50"
[17] "finishedsquarefeet6" "fips" "fireplacecnt" "fullbathcnt"
[21] "garagecarcnt" "garagetotalsqft" "hashottuborspa" "heatingorsystemtypeid"
[25] "latitude" "longitude" "lotssizesquarefeet" "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"
```

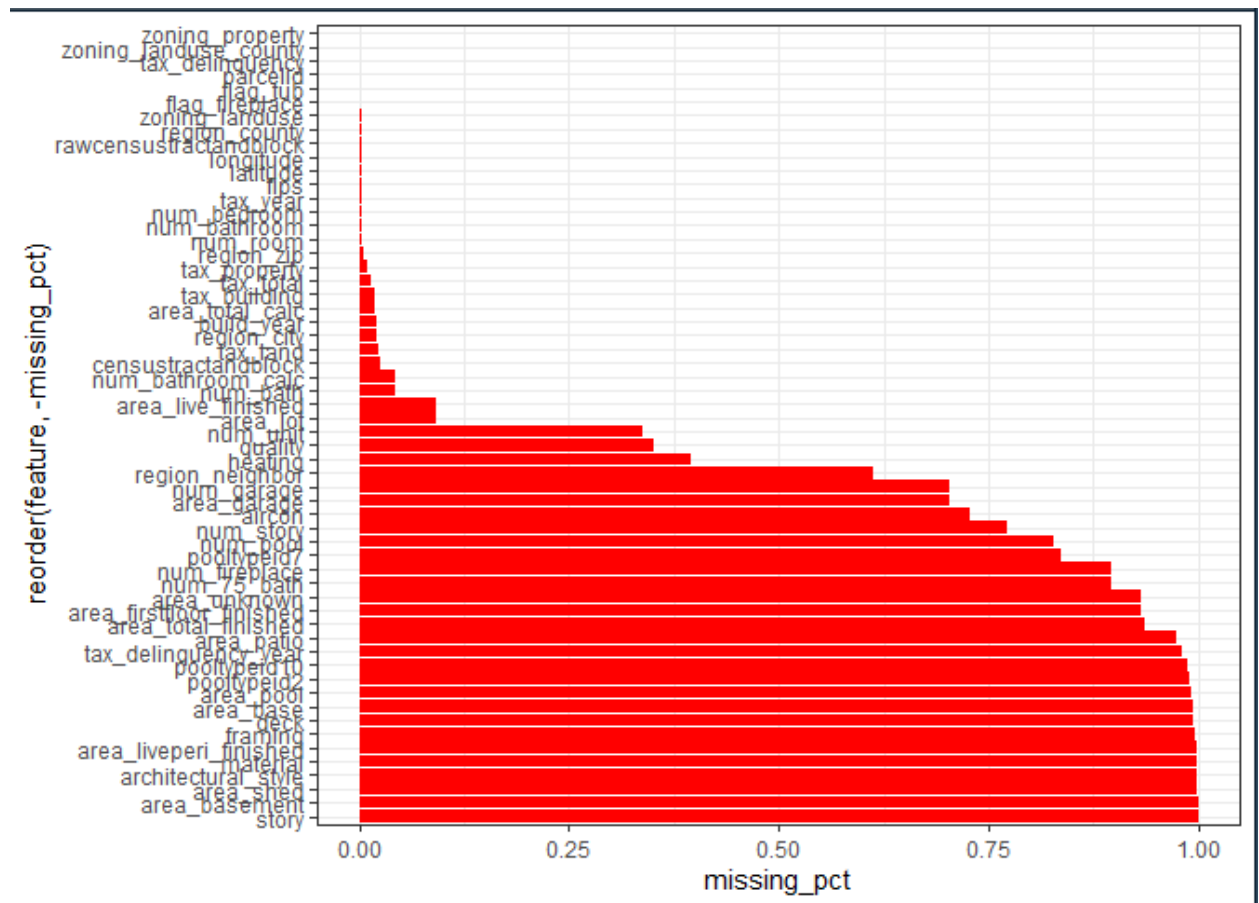
A sample of data is shown below:

File URL:	~/GitHub/Safarie1103/Bellevue University/Courses/DSC520/FinalProject/zipow-prize-1/properties_2016.csv											Update
Data Preview:	lotssizesquarefeet (double)	poolcnt (logical)	poolsizesum (logical)	pooltypeid10 (logical)	pooltypeid2 (logical)	pooltypeid7 (logical)	propertycountylandusecode (character)	propertylandusetypeid (double)	propertyzoningdesc (character)	rawcensustractandblock (character)	regionidcity (double)	regionidcounty (double)
	5333	NA	NA	NA	NA	NA	1210	31	BUC4YY	060373108.003004	396054	3101
	145865	NA	NA	NA	NA	NA	0100	269	BUR1*	060373101.003001	396054	3101
	7494	NA	NA	NA	NA	NA	1210	31	SFC2*	060373202.023005	47547	3101
	3423	NA	NA	NA	NA	NA	1200	47	LAC2	060371112.022021	12447	3101
	81293	NA	NA	NA	NA	NA	0100	269	SCUR3	060379201.081013	NA	3101
	6286	NA	NA	NA	NA	NA	0100	269	LCA25*	060379201.071002	NA	3101
	NA	NA	NA	NA	NA	NA	300V	266	SCBP	060379203.1420	54311	3101
	11975	NA	NA	NA	NA	NA	0100	261	POA1*	060379104.01101	40227	3101
	9403	NA	NA	NA	NA	NA	0100	261	POA21*	060379107.05102	40227	3101
	3817	NA	NA	NA	NA	NA	0100	261	LCA21*	060379107.091112	40227	3101
	8856	NA	NA	NA	NA	NA	1210	31	LAC3*	060379008.061041	5534	3101
	46526	NA	NA	NA	NA	NA	300V	260	LCM*	060379003.001052	5534	3101
	9826	NA	NA	NA	NA	NA	0100	261	LCA22	060379102.061100	40227	3101

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.

```
4 missing_values <- properties %>% summarize_each(funs(sum(is.na(.))/n()))
5 missing_values
6 missing_values <- gather(missing_values, key="feature", value="missing_pct")
7 missing_values %>%
8   ggplot(aes(x=reorder(feature, -missing_pct), y=missing_pct)) +
9     geom_bar(stat="identity", fill="red") +
10    coord_flip() + theme_bw()
11
12
13 good_features <- filter(missing_values, missing_pct < 0.75)
14 |
15 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 than 75% NA's because they would adversely affect our analysis.

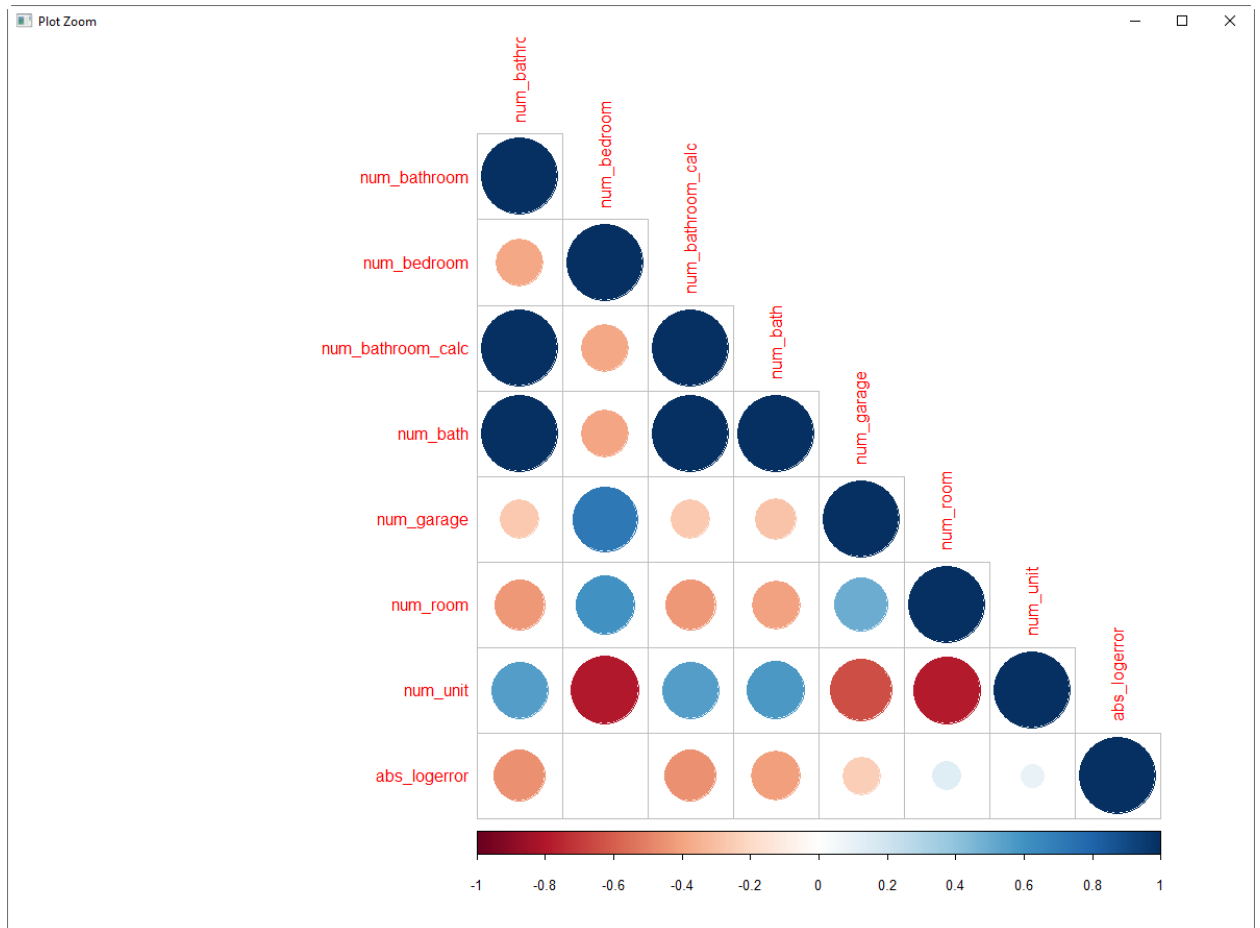
We will use the features in **good_features** dataframe for our analysis. For example, the code below identifies the feature with '_num' in them:

```
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] "parceled" "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

TBD

Section 4 – Final

TBD

Section 1 – Week 9

- Introduction
- Research questions
- Approach
- How your approach addresses (fully or partially) the problem.
- Data
- Required Packages
- Plots and Table Needs
- Questions for future steps.

Section 2 – Week 10

- How to import and clean my data
- What does the final data set look like?
- Questions for future steps.

Section 3 – Week 11

- What information is not self-evident?
- What are different ways you could look at this data?
- How do you plan to slice and dice the data?
- How could you summarize your data to answer key questions?
- What types of plots and tables will help you to illustrate the findings to your questions?
- Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
- Questions for future steps.

Section 4 – Week 12

- A story / narrative that emerged from your data. Follow this structure.
 - Introduction.
 - The problem statement you addressed.
 - How you addressed this problem statement
 - Analysis.
 - Implications.
 - Limitations.
 - Concluding Remarks