

Zillow_Data_Cleaning

Claire Jiaying Wu

10/13/2018

```
library(prettydoc)
library(data.table)
library(DT)
library(stringr)
library(ggpubr)
library(corrplot)
library(caret)
library(RANN)
library(mice)
library(dplyr)
```

Data Cleaning Processes

```
train_raw <- read.csv("../data/train_2016_v2.csv", stringsAsFactors = FALSE)
property <- read.csv("../data/properties_2016.csv", stringsAsFactors = FALSE)
train <- merge(train_raw, property, by = "parcelid")
dim(train)
```

```
[1] 90275    60
```

```
summary(train)
```

```
      parcelid      logerror      transactiondate
Min.   : 10711738  Min.    :-4.60500  Length:90275
1st Qu.: 11559500  1st Qu.: -0.02530  Class :character
Median : 12547337  Median :  0.00600  Mode  :character
Mean   : 12984656  Mean    :  0.01146
3rd Qu.: 14227552  3rd Qu.:  0.03920
Max.   :162960842  Max.    :  4.73700

airconditioningtypeid architecturalstyletypeid basementsqft
Min.   : 1.00      Min.    : 2.00      Min.    : 100.0
1st Qu.: 1.00      1st Qu.: 7.00      1st Qu.: 407.5
Median : 1.00      Median : 7.00      Median : 616.0
Mean   : 1.82      Mean    : 7.23      Mean    : 713.6
3rd Qu.: 1.00      3rd Qu.: 7.00      3rd Qu.: 872.0
Max.   :13.00      Max.    :21.00      Max.    :1555.0
NA's   :61494      NA's    :90014      NA's    :90232

bathroomcnt      bedroomcnt      buildingclasstypeid
Min.   : 0.000     Min.    : 0.000     Min.    : 4
1st Qu.: 2.000     1st Qu.: 2.000     1st Qu.: 4
Median : 2.000     Median : 3.000     Median : 4
Mean   : 2.279     Mean    : 3.032     Mean    : 4
```

3rd Qu.: 3.000	3rd Qu.: 4.000	3rd Qu.:4	
Max. :20.000	Max. :16.000	Max. :4	
	NA's :90259		

buildingqualitytypeid	calculatedbathnbr	decktypeid
Min. : 1.00	Min. : 1.000	Min. :66
1st Qu.: 4.00	1st Qu.: 2.000	1st Qu.:66
Median : 7.00	Median : 2.000	Median :66
Mean : 5.57	Mean : 2.309	Mean :66
3rd Qu.: 7.00	3rd Qu.: 3.000	3rd Qu.:66
Max. :12.00	Max. :20.000	Max. :66
NA's :32911	NA's :1182	NA's :89617

finishedfloor1squarefeet	calculatedfinishedsquarefeet
Min. : 44	Min. : 2
1st Qu.: 938	1st Qu.: 1184
Median :1244	Median : 1540
Mean :1348	Mean : 1773
3rd Qu.:1614	3rd Qu.: 2095
Max. :7625	Max. :22741
NA's :83419	NA's :661

finishedsquarefeet12	finishedsquarefeet13	finishedsquarefeet15
Min. : 2	Min. :1056	Min. : 560
1st Qu.: 1172	1st Qu.:1392	1st Qu.: 1648
Median : 1518	Median :1440	Median : 2104
Mean : 1745	Mean :1405	Mean : 2380
3rd Qu.: 2056	3rd Qu.:1440	3rd Qu.: 2862
Max. :20013	Max. :1584	Max. :22741
NA's :4679	NA's :90242	NA's :86711

finishedsquarefeet50	finishedsquarefeet6	fips	fireplacecnt
Min. : 44	Min. : 257	Min. :6037	Min. :1.00
1st Qu.: 938	1st Qu.:1112	1st Qu.:6037	1st Qu.:1.00
Median :1248	Median :2028	Median :6037	Median :1.00
Mean :1356	Mean :2303	Mean :6049	Mean :1.19
3rd Qu.:1619	3rd Qu.:3431	3rd Qu.:6059	3rd Qu.:1.00
Max. :8352	Max. :7224	Max. :6111	Max. :5.00
NA's :83419	NA's :89854	NA's :80668	

fullbathcnt	garagecarcnt	garagetotalsqft	hashottuborspa
Min. : 1.000	Min. : 0.00	Min. : 0.0	Length:90275
1st Qu.: 2.000	1st Qu.: 2.00	1st Qu.: 0.0	Class :character
Median : 2.000	Median : 2.00	Median : 433.0	Mode :character
Mean : 2.241	Mean : 1.81	Mean : 345.5	
3rd Qu.: 3.000	3rd Qu.: 2.00	3rd Qu.: 484.0	
Max. :20.000	Max. :24.00	Max. :7339.0	
NA's :1182	NA's :60338	NA's :60338	

heatingorsystemtypeid	latitude	longitude
Min. : 1.00	Min. :33339295	Min. :-119447865
1st Qu.: 2.00	1st Qu.:33811538	1st Qu.:-118411692
Median : 2.00	Median :34021500	Median :-118173431
Mean : 3.93	Mean :34005411	Mean :-118198868
3rd Qu.: 7.00	3rd Qu.:34172742	3rd Qu.:-117921588
Max. :24.00	Max. :34816009	Max. :-117554924
NA's :34195		

lotsizesquarefeet	poolcnt	poolsizesum	pooltypeid10
Min. : 167	Min. :1	Min. : 28.0	Min. :1
1st Qu.: 5703	1st Qu.:1	1st Qu.: 420.0	1st Qu.:1
Median : 7200	Median :1	Median : 500.0	Median :1
Mean : 29110	Mean :1	Mean : 519.8	Mean :1
3rd Qu.: 11686	3rd Qu.:1	3rd Qu.: 600.0	3rd Qu.:1
Max. :6971010	Max. :1	Max. :1750.0	Max. :1
NA's :10150	NA's :72374	NA's :89306	NA's :89114

pooltypeid2	pooltypeid7	propertycountylandusecode
-------------	-------------	---------------------------

Min. :1	Min. :1	Length:90275
1st Qu.:1	1st Qu.:1	Class :character
Median :1	Median :1	Mode :character
Mean :1	Mean :1	
3rd Qu.:1	3rd Qu.:1	
Max. :1	Max. :1	
NA's :89071	NA's :73578	

propertylandusetypeid propertyzoningdesc rawcensustractandblock

Min. : 31.0	Length:90275	Min. :60371011
1st Qu.:261.0	Class :character	1st Qu.:60373203
Median :261.0	Mode :character	Median :60376200
Mean :261.8		Mean :60491795
3rd Qu.:266.0		3rd Qu.:60590423
Max. :275.0		Max. :61110091

regionidcity	regionidcounty	regionidneighborhood	regionidzip
Min. : 3491	Min. :1286	Min. : 6952	Min. : 95982
1st Qu.: 12447	1st Qu.:1286	1st Qu.: 46736	1st Qu.: 96193
Median : 25218	Median :3101	Median :118887	Median : 96393
Mean : 33761	Mean :2525	Mean :190646	Mean : 96586
3rd Qu.: 45457	3rd Qu.:3101	3rd Qu.:274800	3rd Qu.: 96987
Max. :396556	Max. :3101	Max. :764167	Max. :399675
NA's :1803	NA's :54263	NA's :35	

roomcnt	storytypeid	threequarterbathnbr
Min. : 0.000	Min. :7	Min. :1.00
1st Qu.: 0.000	1st Qu.:7	1st Qu.:1.00
Median : 0.000	Median :7	Median :1.00
Mean : 1.479	Mean :7	Mean :1.01
3rd Qu.: 0.000	3rd Qu.:7	3rd Qu.:1.00
Max. :18.000	Max. :7	Max. :4.00
	NA's :90232	NA's :78266

typeconstructiontypeid	unitcnt	yardbuildingsqft17
Min. : 4.00	Min. : 1.00	Min. : 25.0
1st Qu.: 6.00	1st Qu.: 1.00	1st Qu.: 180.0
Median : 6.00	Median : 1.00	Median : 259.5
Mean : 6.01	Mean : 1.11	Mean : 310.1
3rd Qu.: 6.00	3rd Qu.: 1.00	3rd Qu.: 384.0
Max. :13.00	Max. :143.00	Max. :2678.0
NA's :89976	NA's :31922	NA's :87629

yardbuildingsqft26	yearbuilt	numberofstories	fireplaceflag
Min. : 18.0	Min. :1885	Min. :1.00	Length:90275
1st Qu.: 100.0	1st Qu.:1953	1st Qu.:1.00	Class :character
Median : 159.0	Median :1970	Median :1.00	Mode :character
Mean : 311.7	Mean :1969	Mean :1.44	
3rd Qu.: 361.0	3rd Qu.:1987	3rd Qu.:2.00	
Max. :1366.0	Max. :2015	Max. :4.00	
NA's :90180	NA's :756	NA's :69705	

structuretaxvaluedollarcnt	taxvaluedollarcnt	assessmentyear
Min. : 100	Min. : 22	Min. :2015
1st Qu.: 81245	1st Qu.: 199023	1st Qu.:2015
Median : 132000	Median : 342872	Median :2015
Mean : 180093	Mean : 457673	Mean :2015
3rd Qu.: 210534	3rd Qu.: 540589	3rd Qu.:2015
Max. :9948100	Max. :27750000	Max. :2015
NA's :380	NA's :1	

landtaxvaluedollarcnt	taxamount	taxdelinquencyflag
Min. : 22	Min. : 49.1	Length:90275
1st Qu.: 82228	1st Qu.: 2872.8	Class :character
Median : 192970	Median : 4542.8	Mode :character
Mean : 278335	Mean : 5984.0	

3rd Qu.:	345420	3rd Qu.:	6901.1
Max.:	24500000	Max.:	321936.1
NA's	:1	NA's	:6
taxdelinquencyyear censustractandblock			
Min.:	6.0	Min.:	6.037e+13
1st Qu.:	13.0	1st Qu.:	6.037e+13
Median:	14.0	Median:	6.038e+13
Mean:	13.4	Mean:	6.049e+13
3rd Qu.:	15.0	3rd Qu.:	6.059e+13
Max.:	99.0	Max.:	6.111e+13
NA's	:88492	NA's	:605

Count the NAs and remove the columns with over 80% NAs

```
# function of counting NAs
count_na = function(x){sum(is.na(x))}
na_count = data.frame(apply(train, 2, count_na))

# computing NA%
na_count$naPercent <- round(na_count[,1]/nrow(train),2)
na_count[order(na_count$naPercent, decreasing = T), ]
```

	apply.train..2..count_na.	naPercent
architecturalstyletypeid	90014	1.00
basementsqft	90232	1.00
buildingclasstypeid	90259	1.00
finishedsquarefeet13	90242	1.00
finishedsquarefeet6	89854	1.00
storytypeid	90232	1.00
typeconstructiontypeid	89976	1.00
yardbuildingsqft26	90180	1.00
decktypeid	89617	0.99
poolsum	89306	0.99
pooltypeid10	89114	0.99
pooltypeid2	89071	0.99
taxdelinquencyyear	88492	0.98
yardbuildingsqft17	87629	0.97
finishedsquarefeet15	86711	0.96
finishedfloor1squarefeet	83419	0.92
finishedsquarefeet50	83419	0.92
fireplacecnt	80668	0.89
threequarterbathnbr	78266	0.87
pooltypeid7	73578	0.82
poolcnt	72374	0.80
numberofstories	69705	0.77
airconditioningtypeid	61494	0.68
garagecarcnt	60338	0.67
garagetotalsqft	60338	0.67
regionidneighborhood	54263	0.60
heatingorsystemtypeid	34195	0.38
buildingqualitytypeid	32911	0.36
unitcnt	31922	0.35
lotsizesquarefeet	10150	0.11
finishedsquarefeet12	4679	0.05
regionidcity	1803	0.02
calculatedbathnbr	1182	0.01
calculatedfinishedsquarefeet	661	0.01
fullbathcnt	1182	0.01
yearbuilt	756	0.01

censustractandblock	605	0.01
parcelid	0	0.00
logerror	0	0.00
transactiondate	0	0.00
bathroomcnt	0	0.00
bedroomcnt	0	0.00
fips	0	0.00
hashottuborspa	0	0.00
latitude	0	0.00
longitude	0	0.00
propertycountylandusecode	0	0.00
propertylandusetypeid	0	0.00
propertyzoningdesc	0	0.00
rawcensustractandblock	0	0.00
regionidcounty	0	0.00
regionidzip	35	0.00
roomcnt	0	0.00
fireplaceflag	0	0.00
structuretaxvaluedollarcnt	380	0.00
taxvaluedollarcnt	1	0.00
assessmentyear	0	0.00
landtaxvaluedollarcnt	1	0.00
taxamount	6	0.00
taxdelinquencyflag	0	0.00

```
# excluding the features that have too many missing values
KeepCol <- rownames(na_count[na_count$naPercent <= 0.2,])
train_sub <- train[, KeepCol]
```

```
dim(train_sub)
```

```
[1] 90275    31
```

```
str(train_sub)
```

```
'data.frame':  90275 obs. of  31 variables:
 $ parcelid      : int  10711738 10711755 10711805 10711816 10711858 10711910
 $ logerror      : num  0.0276 -0.0182 -0.1009 -0.0121 -0.0481 ...
 $ transactiondate : chr  "2016-08-02" "2016-08-02" "2016-05-03" "2016-04-05" ..
 $ bathroomcnt   : num  3 3 2 2 2 2 2 3 3 3 ...
 $ bedroomcnt    : num  4 3 3 4 4 3 4 3 4 3 ...
 $ calculatedbathnbr : num  3 3 2 2 2 2 2 3 3 3 ...
 $ calculatedfinishedsquarefeet: num  2538 1589 2411 2232 1882 ...
 $ finishedsquarefeet12 : int  2538 1589 2411 2232 1882 1477 1850 3193 2421 1678 ...
 $ fips          : int  6037 6037 6037 6037 6037 6037 6037 6037 6037 6037 ...
 $ fullbathcnt    : int  3 3 2 2 2 2 2 3 3 3 ...
 $ hashottuborspa : chr  "" "" "" "" ...
 $ latitude       : int  34220381 34222040 34220427 34222390 34222544 34221864
 $ longitude      : int  -118620802 -118622240 -118618549 -118618631 -118617961
 $ lotsizesquarefeet : num  11012 11010 11723 9002 9002 ...
 $ propertycountylandusecode : chr  "0101" "0101" "0101" "0100" ...
 $ propertylandusetypeid : int  261 261 261 261 261 261 261 261 261 261 ...
 $ propertyzoningdesc : chr  "LARE11" "LARE11" "LARE9" "LARE9" ...
 $ rawcensustractandblock : num  60371132 60371132 60371132 60371132 60371132 ...
 $ regionidcity    : int  12447 12447 12447 12447 12447 12447 12447 12447 12447
 $ regionidcounty  : int  3101 3101 3101 3101 3101 3101 3101 3101 3101 3101 ...
 $ regionidzip     : int  96339 96339 96339 96339 96339 96339 96339 96339 96339
```

```

$ roomcnt           : num  0 0 0 0 0 0 0 0 0 ...
$ yearbuilt         : num  1978 1959 1973 1973 1973 ...
$ fireplaceflag     : chr   "" "" "" "" ...
$ structuretaxvaluedollarcnt : num  245180 254691 235114 262309 232037 ...
$ taxvaluedollarcnt : num  567112 459844 384787 437176 382055 ...
$ assessmentyear    : int   2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
$ landtaxvaluedollarcnt : num  321932 205153 149673 174867 150018 ...
$ taxamount         : num  7219 6901 4877 5560 4878 ...
$ taxdelinquencyflag : chr   "" "" "" "" ...
$ censustractandblock : num  6.04e+13 6.04e+13 6.04e+13 6.04e+13 6.04e+13 ...

```

Coerce the variables into right data types

```

char_cols = c('fips', 'propertylandusetypeid', 'rawcensustractandblock', 'regionidcounty',
              'assessmentyear', 'regionidzip', 'censustractandblock', 'regionidcity')
train_sub[,char_cols] = apply(train_sub[,char_cols], 2, function(x) as.character(x))
train_sub$taxdelinquencyflag = ifelse(train_sub$taxdelinquencyflag != "", TRUE, FALSE)
bool_cols = c("hashottuborspa", "fireplaceflag")
train_sub[,bool_cols] <- apply(train_sub[,bool_cols], 2, function(x) as.logical(x))
train_sub$yearbuilt = as.character(train_sub$yearbuilt)

str(train_sub)

```

```

'data.frame':  90275 obs. of  31 variables:
 $ parcelid           : int  10711738 10711755 10711805 10711816 10711858 10711910
 $ logerror           : num  0.0276 -0.0182 -0.1009 -0.0121 -0.0481 ...
 $ transactiondate    : chr   "2016-08-02" "2016-08-02" "2016-05-03" "2016-04-05" ..
 $ bathroomcnt       : num  3 3 2 2 2 2 2 3 3 3 ...
 $ bedroomcnt        : num  4 3 3 4 4 3 4 3 4 3 ...
 $ calculatedbathnbr  : num  3 3 2 2 2 2 2 3 3 3 ...
 $ calculatedfinishedsquarefeet: num  2538 1589 2411 2232 1882 ...
 $ finishedsquarefeet12 : int  2538 1589 2411 2232 1882 1477 1850 3193 2421 1678 ...
 $ fips              : chr   "6037" "6037" "6037" "6037" ...
 $ fullbathcnt       : int  3 3 2 2 2 2 2 3 3 3 ...
 $ hashottuborspa    : logi  NA NA NA NA NA NA ...
 $ latitude          : int  34220381 34222040 34220427 34222390 34222544 34221864
 $ longitude         : int  -118620802 -118622240 -118618549 -118618631 -118617961
 $ lotsizesquarefeet : num  11012 11010 11723 9002 9002 ...
 $ propertycountylandusecode : chr   "0101" "0101" "0101" "0100" ...
 $ propertylandusetypeid : chr   "261" "261" "261" "261" ...
 $ propertyzoningdesc  : chr   "LARE11" "LARE11" "LARE9" "LARE9" ...
 $ rawcensustractandblock : chr   "60371132.32102" "60371132.321019" "60371132.32102" "60371132.32102" ...
 $ regionidcity       : chr   "12447" "12447" "12447" "12447" ...
 $ regionidcounty     : chr   "3101" "3101" "3101" "3101" ...
 $ regionidzip        : chr   "96339" "96339" "96339" "96339" ...
 $ roomcnt           : num  0 0 0 0 0 0 0 0 0 ...
 $ yearbuilt         : chr   "1978" "1959" "1973" "1973" ...
 $ fireplaceflag     : logi  NA NA NA NA NA NA ...
 $ structuretaxvaluedollarcnt : num  245180 254691 235114 262309 232037 ...
 $ taxvaluedollarcnt : num  567112 459844 384787 437176 382055 ...
 $ assessmentyear    : chr   "2015" "2015" "2015" "2015" ...
 $ landtaxvaluedollarcnt : num  321932 205153 149673 174867 150018 ...
 $ taxamount         : num  7219 6901 4877 5560 4878 ...
 $ taxdelinquencyflag : logi  FALSE FALSE FALSE FALSE FALSE ...
 $ censustractandblock : chr   "60371132321020" "60371132321019" "60371132321020" "60371132321020" ...

```

Impute missing values with simple imputation

```

beMedian <- function(x){ifelse(is.na(x), median(x, na.rm = T), x)}
beMean <- function(x){ifelse(is.na(x), mean(x, na.rm = T), x)}
beMode <- function (x) {
  xtab <- table(x)
  ifelse(is.na(x), names(which(xtab == max(xtab))), x)
}
beOpposite <- function(x){ifelse(is.na(x), FALSE, x)}

missing_impute <- function(dataframe){
  col_class <- sapply(dataframe, class)
  col_num <- names(col_class[col_class %in% c("numeric", "integer")])
  col_cat <- names(col_class[col_class %in% c("character", "factor")])
  col_log <- names(col_class[col_class %in% c("logical")])

  dataframe[, col_num] <- apply(dataframe[, col_num], 2, beMedian)
  dataframe[, col_cat] <- apply(dataframe[, col_cat], 2, beMode)
  dataframe[, col_log] <- apply(dataframe[, col_log], 2, beOpposite)
  return(cbind(dataframe[, col_num], dataframe[, col_cat], dataframe[, col_log]))
}

train_imputed <- missing_impute(train_sub)
train_imputed[is.na(train_imputed)]

```

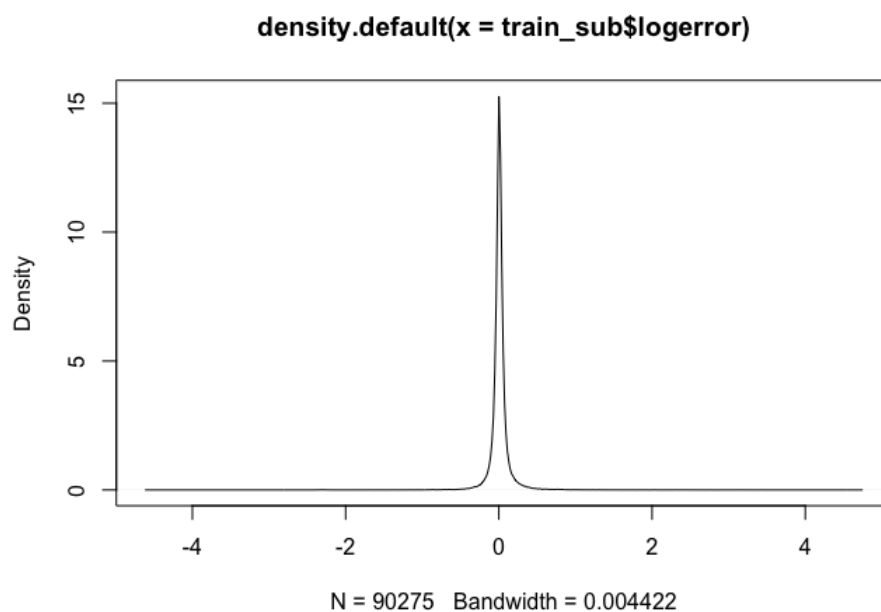
```
character(0)
```

```
write.csv(train_imputed, file = "../cleaning/property_cleaned.csv", row.names = F)
```

EDA

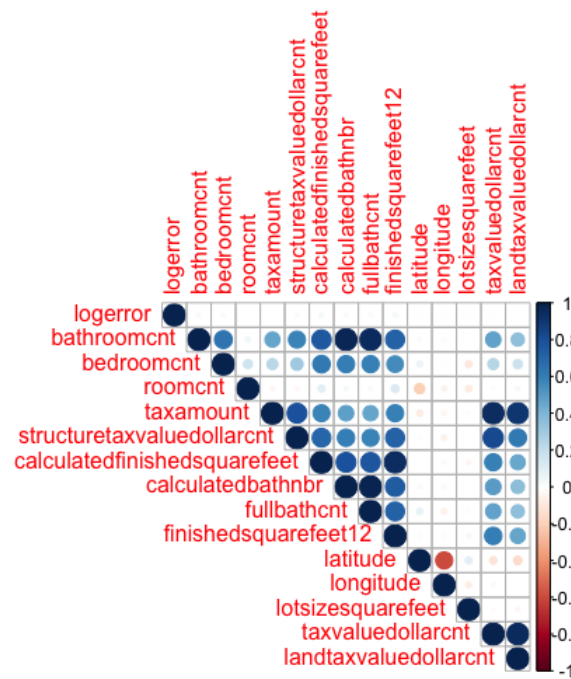
Density plot for the numeric variable with kernel smoothing:

```
plot(density(train_sub$logerror))
```



Explore correlations between numerical variables and output with corplot:

```
correlations <- cor(train_imputed[, c('logerror', 'bathroomcnt', 'bedroomcnt', 'roomcnt', 'taxamount', 'structuretaxvaluedollarcnt', 'calculatedfinishedsquarefeet', 'calculatedbathnbr', 'fullbathcnt', 'finishedsquarefeet12', 'latitude', 'longitude', 'lotsizesquarefeet', 'taxvaluedollarcnt', 'landtaxvaluedollarcnt')])
corrplot(correlations, method = "circle", type = 'upper')
```



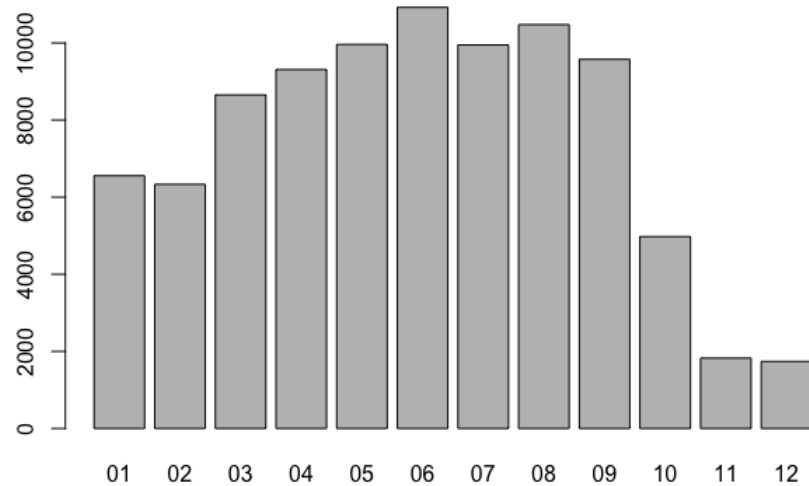
No numeric independent variable has high correlation with the output.

Explore the categorical variables (by month):

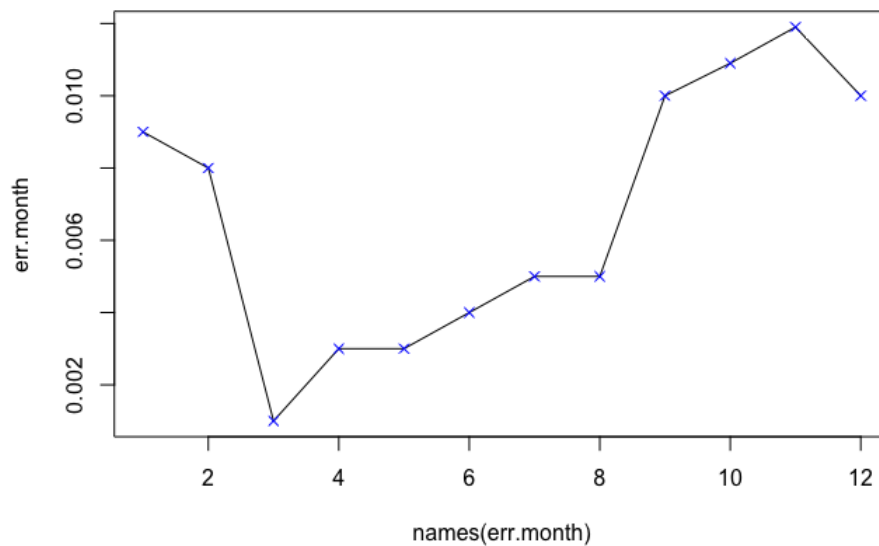
```
# extract month from transaction date
train_sub$txnmonth <- format(as.Date(train_sub$transactiondate), "%m")
table(train_sub$txnmonth)
```

01	02	03	04	05	06	07	08	09	10	11	12
6556	6333	8652	9311	9961	10922	9947	10476	9575	4977	1826	1739

```
barplot(table(train_sub$txnmonth))
```

```
# compute the median of log error by month and visualize the points
err.month <- by(train_sub, train_sub$txnmonth, function(x) {return(median(x$logerror))})
plot(names(err.month), err.month, type = 'l')
points(err.month, pch = 4, col = "blue")
```



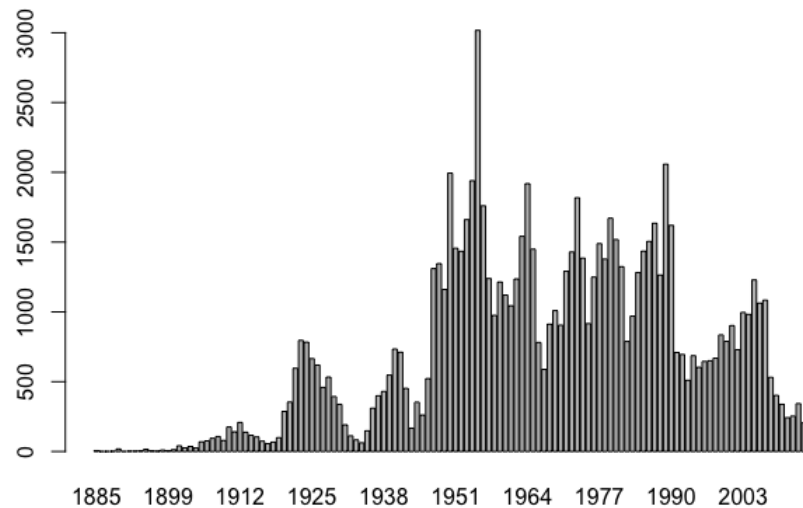
Explore the categorical variables (by yearbuilt):

```
table(train_imputed$yearbuilt)
```

```
1885 1886 1887 1888 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 1900
  6    1    1    2   18    1    3    3    5   16    5    3   10    6   13
1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915
 41   26   36   25   69   76   94  107   79  175  141  208  138  116  104
```

1916	1917	1918	1919	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930
74	56	67	99	287	354	596	796	783	665	619	459	533	393	338
1931	1932	1933	1934	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944	1945
191	112	83	62	148	310	399	430	547	734	711	452	167	353	261
1946	1947	1948	1949	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959	1960
522	1310	1346	1161	1994	1456	1434	1661	1940	3017	1760	1239	975	1214	1120
1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975
1044	1236	1541	1919	1449	781	588	912	1009	905	1291	1429	1818	1385	916
1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
1249	1489	1378	1671	1518	1324	788	969	1282	1435	1505	1635	1264	2058	1620
1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005
709	695	510	686	603	645	650	668	835	789	902	729	996	982	1229
2006	2007	2008	2009	2010	2011	2012	2013	2014	2015					
1063	1084	531	402	338	242	253	343	204	23					

```
barplot(table(train_imputed$yearbuilt))
```

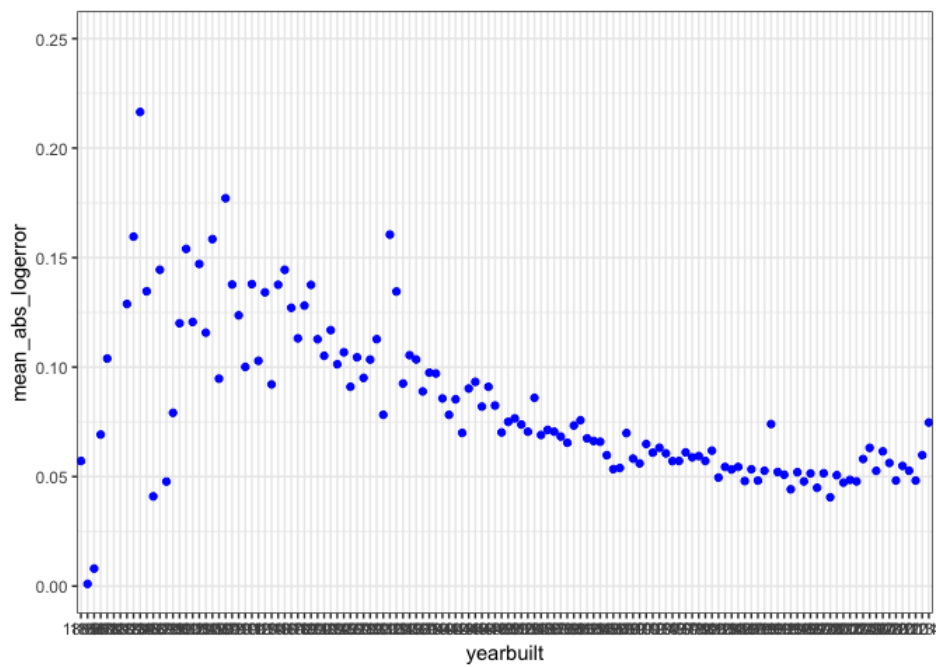


More targets were built

during 1949-1991, based on the barplot. Consider to dip deep.

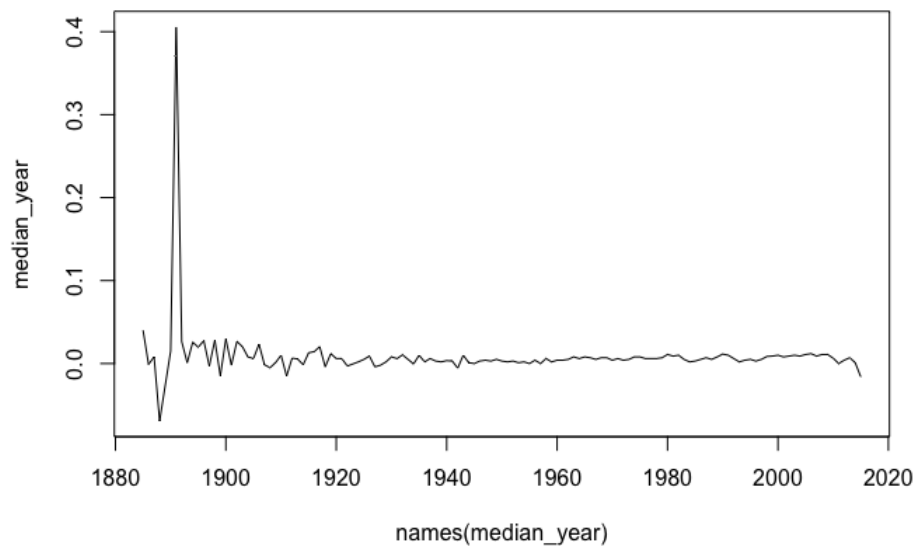
Review the logerror mean distribution:

```
train_imputed %>%
  group_by(yearbuilt) %>%
  summarize(mean_abs_logerror = mean(abs(logerror)), n()) %>%
  ggplot(aes(x=yearbuilt, y=mean_abs_logerror)) +
  geom_smooth(color="grey40") +
  geom_point(color="blue") + coord_cartesian(ylim=c(0, 0.25)) + theme_bw()
```



Check the median of logerror across years:

```
median_year <- by(train_imputed, train_imputed$yearbuilt, function(x) {return(median(x$logerror))})
plot(names(median_year), median_year, type = 'l')
```



Here, we can tell from the plot that one of the features could be yearbuilt, meaning for the very old houses, the logerror tends to be higher than those relatively new houses.