Data exploration for modeling property price

LYT

2019.3.21

Update 25 March

Talked to Chief editor of International Journal of Data Analytics: Web of Science, Scopus, Inspec, PsycINFO, Ei Compendex, and so on

Issue

Sparse time-series transaction Data size and computation speed

Project objective

price estimation, updating past property worth records in the dataset to the present value

I will try to build a model that is able to predict current prices of houses in record, based on their attribute, spatial information and historical transactions to reflect the market value of them. This helps to capture the economic risk of tidal flooding in a more intuitive way.

Timeline

late Mar: a trial machine learning algorithm (Langyi more on this part)

May: a good visualization (also for class project, Bin and Julia will be more involved here)

Methodology

Can try simple tree, random forest, and boosting) first for now. Random variable selection limitation can help to reduce correlated features issue. Can have variable importance measure. More robust model.

No dimentionality reduction for now, cuz there are not so many variables yet.

Will use package randomForest and caret.

Feed multiple models across parcels. Consider ZIP/county/metrolevel or other indicators of neighbourhood. But whether to use the same group of variables can be an issue since some variables miss by county.

How to deal with variables that change over time? Variable of time can be put into model itself, but introducing others will bring covariates.

If use CV: training vs. testset, how to divide? Across parcels? Look at descriptive statistics across counties to think about heterogenity (there's difference between small/big/rural/urban counties). What about 10-fold or LOOCA?

litrature review:

In terms of property valuation, mainstream method is parametric hedonic regression. Machine learning came into application recently. 2 papers are the most relevant for now.

Barr et al. (2017) used gradient boosting trees (offers some interpretability) to estimate individual home price at each periods to constuct a house price index. They suggested that local aggregation (metro, county, state, etc.) is more appropriate than global aggregation, as local trends depart from general trend from time to time. They raised the idea of “submarket” as cohort of houses that competing for the same group of people. Therefore, they run many millions of models across geographic hierarchies (but didn’t say more specifically). They didn’t mention the variables they are using and whether they perform data reduction though.

Garcia-Magarino et al. (2019) tested several machine learning and dimensionality reduction methods to address the problem of estimating the missing prices of a sample of houses. They tried OLS, KNN, SVR (an adaptation of SVM), and Artificial neural networks. Dimensionality reduction methods included Non-negative Matrix Factorization, Recursive Feature Elimination, and Forward Selection.

Data description

There are three sets of data records utilized in the project:

1. home attribute data (codebook: 5.0 Tax Assessor Layout)
2. sales records/transactions (codebook: 5.0 Recorder Layout)
3. flooding risk / environmental variables parcel level variables (parcel risk and spatial data) and one set of polygon parcel boundaries for most of the parcels in these three datasets (sef_parcels.zip)

The parcel attribute and sales data both have an identifier (attomid) for each property. The parcel risk / spatial data file can be joined to this data as it also contains attomid. The parcel polygons and parcel risk / spatial data both have another id (fsid / firststreetid) that can be used to combine each unique parcel.

The parcel risk / spatial data file contains fields that represent the inundation risk with field lengths of 6 or 8, e.g. ltc118, rdkt27, mdc118. The first two characters (lt, rd, md, np) represent whether the statistic is about the proportion of the lot, the proportion of roads nearby, the max depth of inundation on the lot (ceiled to feet), or the proportion of nearby properties impacted. The next two characters (kt, em, c1, c3, c5) represent the risk type: kt for repeated king tides, em for highest annual tide, and c1, c3, c5 for hurricane types. The next two characters represent the year for the risk, 18 for 2018, 23 for 2023, etc. If you find te or qu as characters 7 and 8, it identifies the spatial radius used for the measure, tenth of a mile or a quarter mile.

Considering the input variables, tax data can be used to adjust for market price in cases of missing information; the home attributes data vary by county so we should consider hierarchical modelling, if the trial model reveals significance of the unique variables; need to find a proper way to aggregate environment data.

The home characteristics data

In local desktop I only imported 10000 obs for trial.

```
library(tidyr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
home_dta_original<-fread("D:/raw_data/SF_Home_Characteristics.csv")
```

Brief summary

```
cat("Data include state:",
    unique(home_dta_original$situsstatecode)%>%as.character(),
    ",and county:",
    unique(home_dta_original$situscounty)%>%as.character(),
    "In each state the number of samples are"
    )
```

```
## Data include state: FL ,and county: Broward Miami-Dade Palm Beach Collier Lee Monroe Hendry In each
state the number of samples are
```

```
group_by(home_dta_original,situscounty)%>%summarise(n(),
                                          mean(assessorlastsaleamount,na.rm = TRUE),
                                          mean(areabuilding,na.rm = TRUE)
                                          )
```

```
## # A tibble: 7 x 4
##   situscounty  `n()` `mean(assessorlastsaleamount~ `mean(areabuilding, na.~
##   <chr>        <int>                         <dbl>                    <dbl>
## 1 Broward     759392                       344927.                    2207.
## 2 Collier     290821                       886415.                    1500.
## 3 Hendry       35908                       576729.                     814.
## 4 Lee         574084                       347050.                    1390.
## 5 Miami-Dade  931150                       441260.                    2406.
## 6 Monroe       91360                       316645.                    1007.
## 7 Palm Beach  645208                       492636.                    2237.
```

Considerable level of heterogenity by county might exist in data.

Select and recode useful variables

```
home_dta<-select(home_dta_original,
                 attomid,
                 deedlastsaleprice,
                 situsstatecode,
                 situscounty,
                 ownertypedescription1,
                 ownertypedescription2,
                 yearbuilt,
                 propertyusegroup,
                 deedlastsaledate,
                 areabuilding,
                 roomsatticflag,
                 parkinggarage:communityrecroomflag)
#Fill in the price variable so that it will not be dropped later
home_dta$deedlastsaleprice[is.na(home_dta$deedlastsaleprice)=="TRUE"]<-0
#Make id numeric
home_dta$attomid<-home_dta$attomid%>%as.numeric()
#Owner type recoding misseallenous to NA
home_dta$ownertypedescription1[home_dta$ownertypedescription1=="NP"]<-NA
home_dta$ownertypedescription1[home_dta$ownertypedescription1=="UNKNOWN"]<-NA
home_dta$ownertypedescription2[home_dta$ownertypedescription1=="NP"]<-NA
home_dta$ownertypedescription2[home_dta$ownertypedescription1=="UNKNOWN"]<-NA
#Recoding property use group
home_dta$propertyusegroup[home_dta$propertyusegroup=="UNKNOWN"|
                            home_dta$propertyusegroup=="Other"|
                            home_dta$propertyusegroup=="NP"]<-NA
#152 PropertyUseStandardized is better coded by
class_coding<-read.csv("D:/raw_data/prop_use_codes_trim.csv")
#Rounding deed last sale date to year and recoding NAs
library(stringr)
home_dta$deedlastsaledate<-str_sub(home_dta$deedlastsaledate,start = 0,end = 4)%>%
  as.numeric()
home_dta$deedlastsaledate[home_dta$deedlastsaledate==""]<-NA
#Excluding <50 sq. feet living area
home_dta$areabuilding[home_dta$areabuilding<50]<-NA
#Recoding parkinggarage (?)
home_dta$parkinggarage[home_dta$parkinggarage=="11"|
                                                  home_dta$parkinggarage=="12"|
                                                  home_dta$parkinggarage=="18"|
                                                  home_dta$parkinggarage=="40"|
                                                  home_dta$parkinggarage=="999"]<-NA
#Other variables from parkinggarage yet to recode
```

```
#Some rough recodings to get rid of character
for (i in 1:ncol(home_dta)){
  if (class(home_dta[[i]])=="character"){
    print(names(home_dta)[i])
  }
}
```

```
## [1] "situsstatecode"
## [1] "situscounty"
## [1] "ownertypedescription1"
## [1] "ownertypedescription2"
## [1] "propertyusegroup"
## [1] "exterior1code"
## [1] "viewdescription"
## [1] "porchcode"
```

```
home_dta$situsstatecode<-home_dta$situsstatecode%>%as.factor
home_dta$situscounty<-home_dta$situscounty%>%as.factor
home_dta$ownertypedescription1<-home_dta$ownertypedescription1%>%as.factor
home_dta$ownertypedescription2<-home_dta$ownertypedescription2%>%as.factor
home_dta$propertyusegroup<-home_dta$propertyusegroup%>%as.factor
home_dta$viewdescription<-home_dta$viewdescription%>%as.factor()
home_dta$porchcode<-home_dta$porchcode%>%as.factor()
#Delete exterior code due to too many factor levels
home_dta$exterior1code<-NULL
```

Modeling by county

```
home_county_dta<-split(home_dta,home_dta$situscounty)
```

```
#Function to determine whether a variable is missing less than 10% values
is.missing<-function(x){
    a<-x%>%length()
    b<-x%>%is.na()%>%sum()
    if (b/a<0.1){
      return(TRUE)
    }
    else{
      return(FALSE)
    }
  }

#Function to preprocess a data frame and drop according to is.missing
drop.missing<-function(x){
  for (i in names(x)){
    if (is.missing(x[[i]])==FALSE){
      x<-select(x,-i)
    }
  }
  x<-x
}

#Process data by county
library(purrr)
```

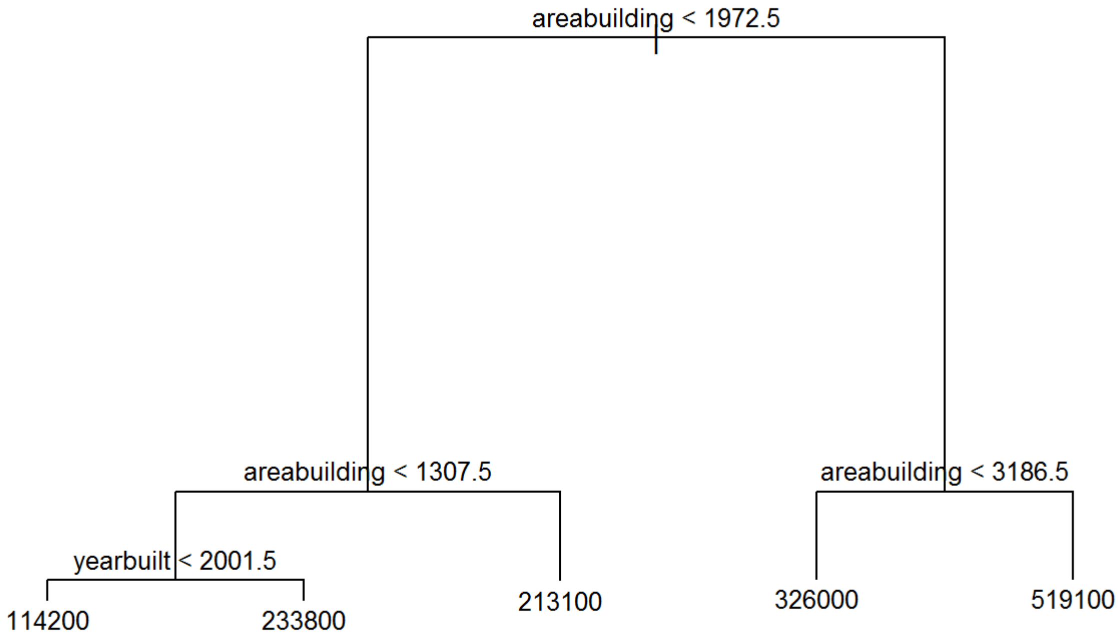
```
##
## Attaching package: 'purrr'
```

```
## The following object is masked from 'package:data.table':
##
##     transpose
```

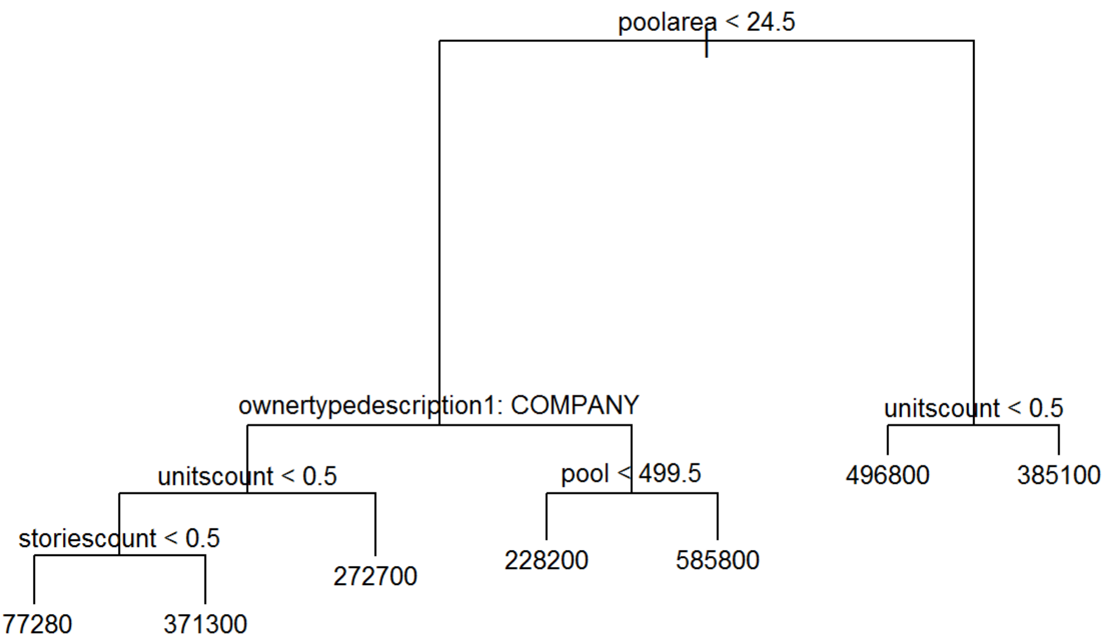
```
home_county_cleaned_dta<-map(home_county_dta,drop.missing)
```

```
#Try simple tree
tree.model<-function(dta){
#Report data subset
cat("Below prints model for this state:",
    dta$situsstatecode%>%unique()%>%as.character(),
    "and this county:",
    dta$situscounty%>%unique()%>%as.character()
)
#Excluding<$1000 transactions which are not authentic
dta$deedlastsaleprice[dta$deedlastsaleprice<1000]<-NA
dta$deedlastsaleprice<-dta$deedlastsaleprice%>%
  as.numeric()
#Split training/test samples (0.7:0.3)
train<-sample_frac(dta,size=0.7)
test<-anti_join(dta,train,by="attomid")
#Missing value check (unused here)
check<-function(train){
for (i in 1:ncol(train)){
      a<-train[i]%>%nrow()
    b<-train[i]%>%is.na()%>%sum()
    c<-b/a
    print(c)
}
}
#Prepare y and x features
train<-select(train,
          -attomid,
          -situsstatecode,
          -situscounty)
y<-train$deedlastsaleprice
x<-select(train,
          -deedlastsaleprice)
#Simple tree
library(tree)
#Excluding>$1000000 transactions which are not authentic
y[y>1000000]<-NA
#Simple tree
train_tree<-tree(y~.,x,
                 na.action="na.omit")
plot(train_tree)
text(train_tree, pretty = 0, cex = .8)
}
#tree.model(home_county_cleaned_dta$Hendry)
map(home_county_cleaned_dta,tree.model)
```

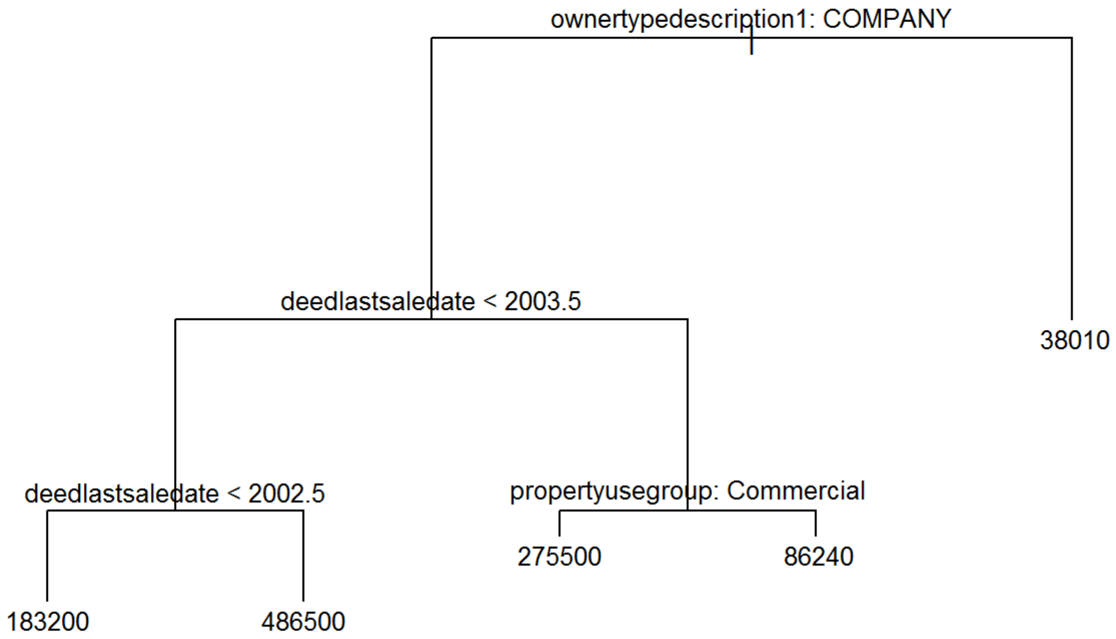
```
## Below prints model for this state: FL and this county: Broward
```



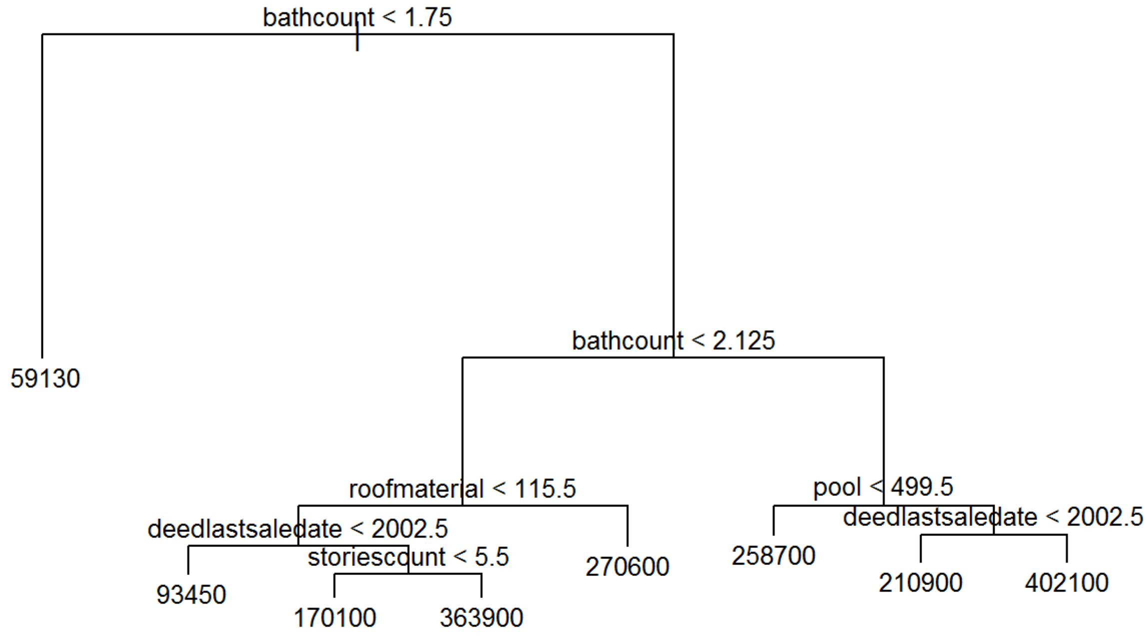

```
## Below prints model for this state: FL and this county: Collier
```



```
## Below prints model for this state: FL and this county: Hendry
```

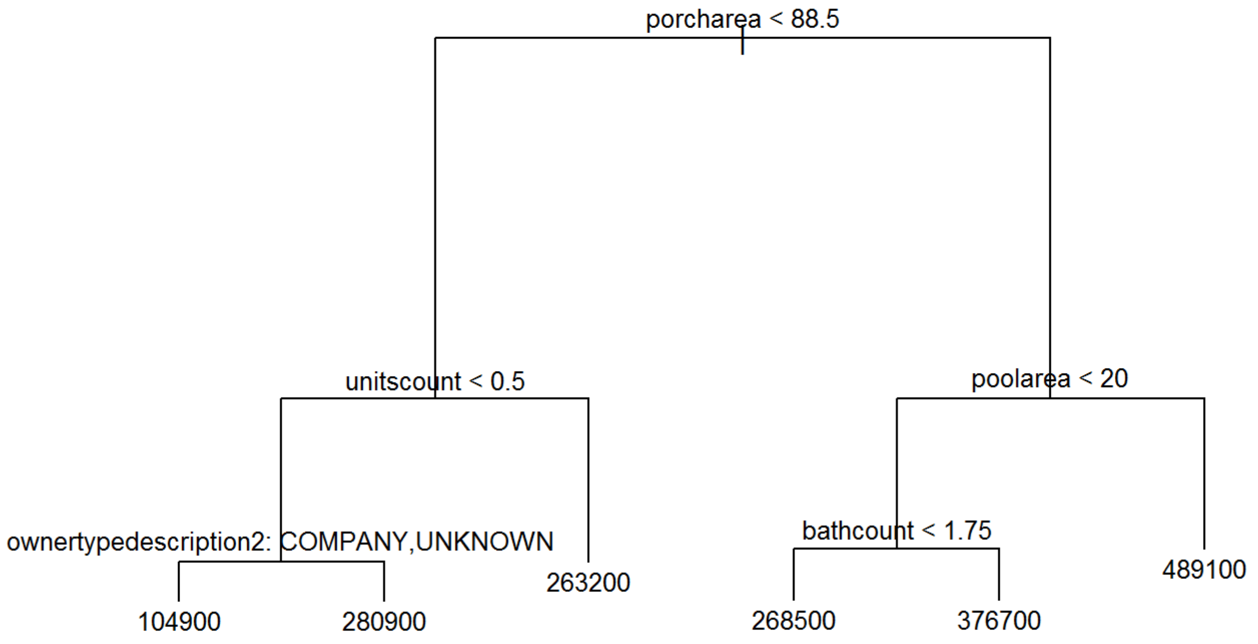



```
## Below prints model for this state: FL and this county: Lee
```

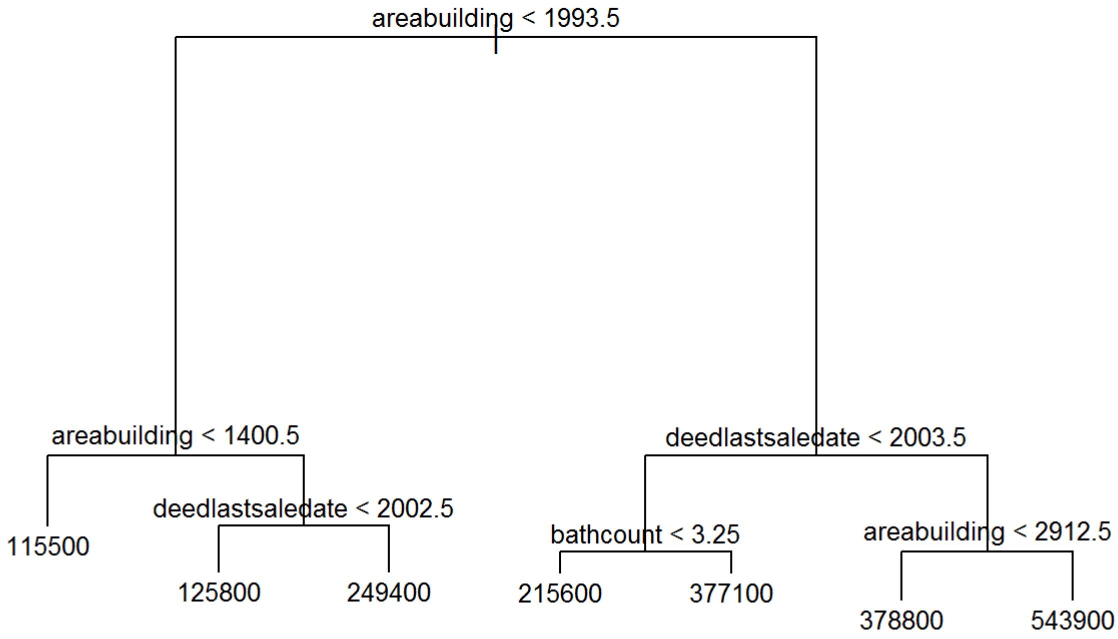



```
## Below prints model for this state: FL and this county: Miami-Dade
```

```
## Below prints model for this state: FL and this county: Monroe
```




```
## Below prints model for this state: FL and this county: Palm Beach
```




```
## $Broward
## NULL
##
## $Collier
## NULL
##
## $Hendry
## NULL
##
## $Lee
## NULL
##
## $`Miami-Dade`
## NULL
##
## $Monroe
## NULL
##
## $`Palm Beach`
## NULL
```

```
#Because there are few nodes, I don't make prediction on test set here.
```

```
#Try random forest
rf.model<-function(dta){
#Report data subset
cat("Below prints model for this state:",
    dta$situsstatecode%>%unique()%>%as.character(),
    "and this county:",
    dta$situscounty%>%unique()%>%as.character()
)
#Reduce size (30000) for computation convenience
  if (nrow(dta)>30000){
    dta<-sample_n(dta,30000)
  }
#Excluding>$500000 transactions (extreme values) to experiment
dta$deedlastsaleprice[dta$deedlastsaleprice>50000]<-NA
#Excluding<$1000 transactions which are not authentic
dta$deedlastsaleprice[dta$deedlastsaleprice<1000]<-NA
dta$deedlastsaleprice<-dta$deedlastsaleprice%>%
  as.numeric()
#Split training/test samples (0.7:0.3)
train<-sample_frac(dta,size=0.7)
test<-anti_join(dta,train,by="attomid")
#Missing value check (unused here)
check<-function(train){
for (i in 1:ncol(train)){
      a<-train[i]%>%nrow()
    b<-train[i]%>%is.na()%>%sum()
    c<-b/a
    print(c)
}
}
#Prepare y and x features
y<-train$deedlastsaleprice
x<-select(train,
          -attomid,
          -deedlastsaleprice,
          -situsstatecode,
          -situscounty)
#Random forest
library(randomForest)
train_rf <- randomForest(y~.,x,
                         importance = TRUE,
                         na.action = "na.omit"
                         )
importance(train_rf)%>%print()
#Test set
yhat.rf <- predict(train_rf,test)
cat("The test MSE is",
    mean((yhat.rf-test$deedlastsaleprice)^2,na.rm = TRUE)
    )
plot(yhat.rf,test$deedlastsaleprice,
     cex = .2)%>%print()
abline(0,1)
}
#rf.model(home_county_cleaned_dta$Hendry)
map(home_county_cleaned_dta,rf.model)
```

```
## Below prints model for this state: FL and this county: Broward
```

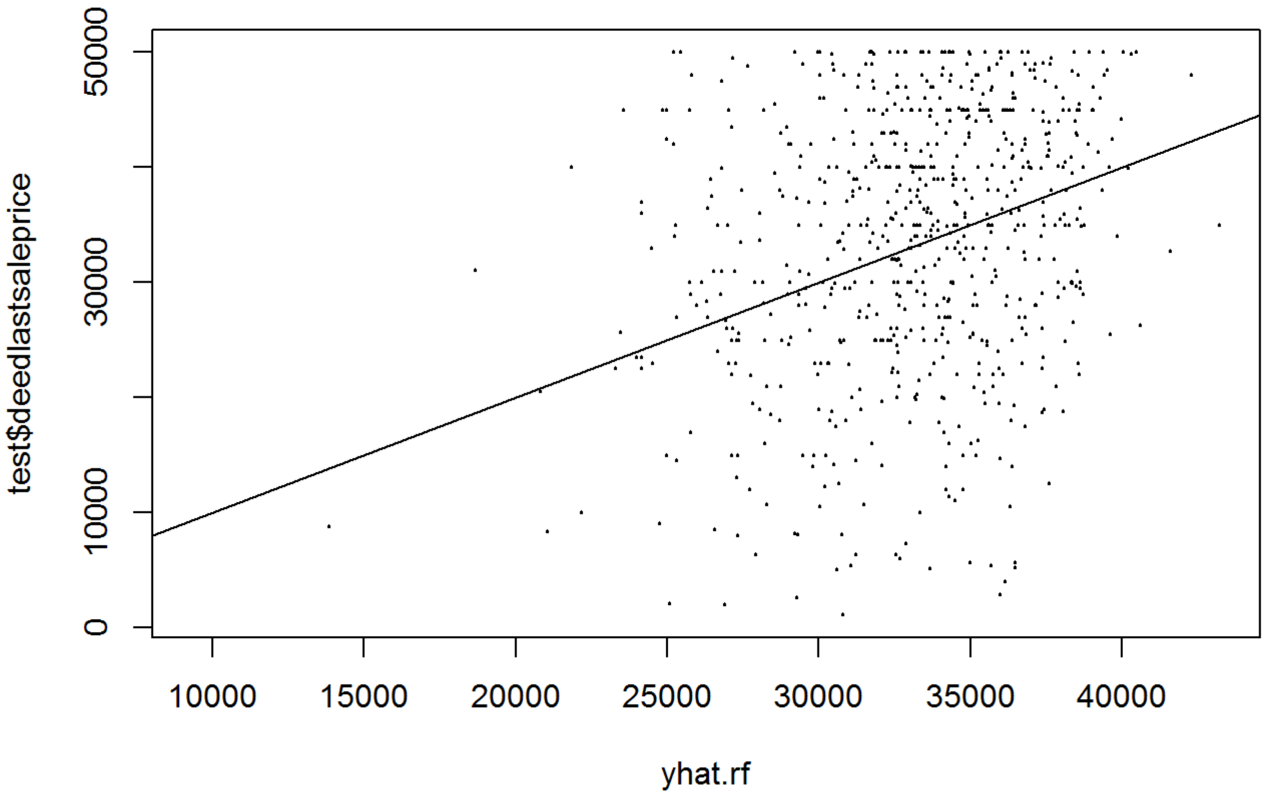
```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

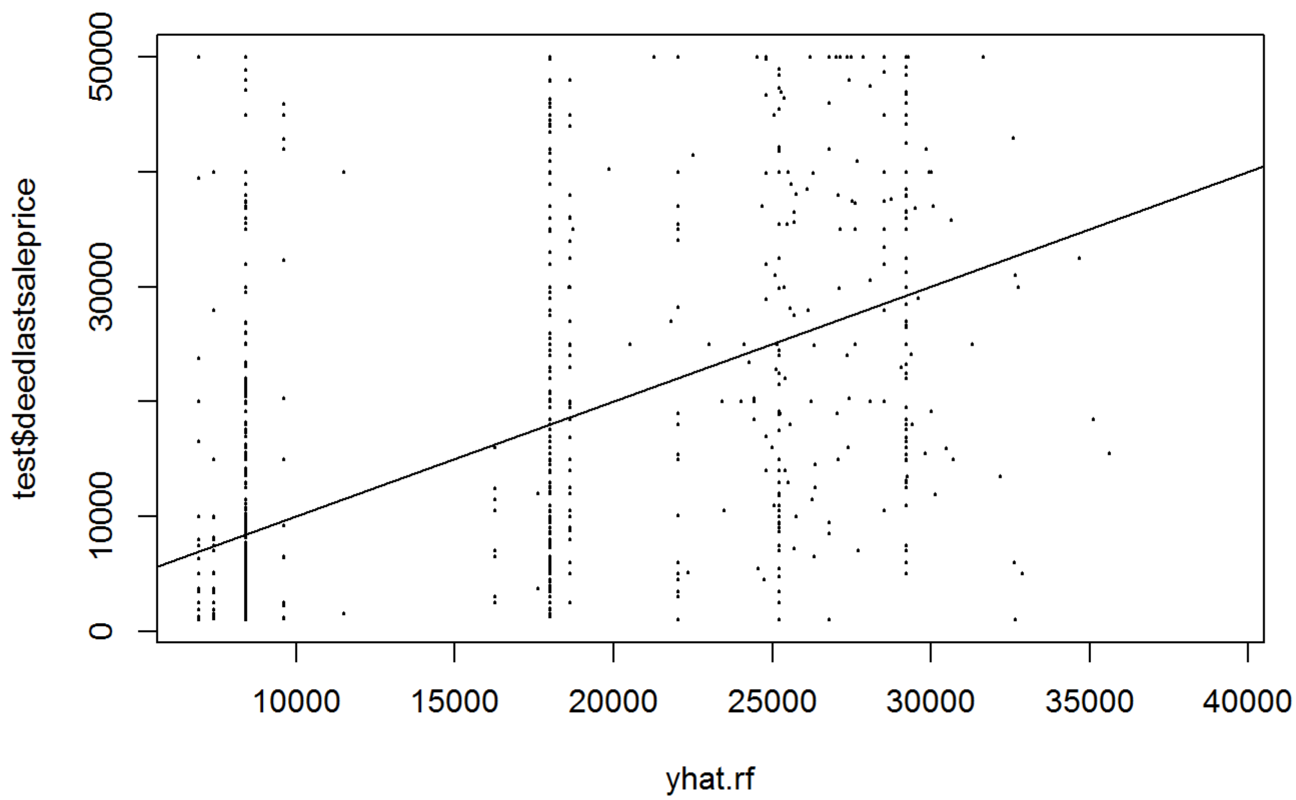
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
##                          %IncMSE IncNodePurity
## ownertypedescription1 16.5649741    4490921480
## ownertypedescription2  1.9299381    6309763211
## yearbuilt             25.1940348   28630022115
## propertyusegroup       5.1564927    1721988855
## areabuilding          25.3626146   40227436972
## parkinggarage          0.0000000             0
## parkinggaragearea      0.0000000             0
## hvacheatingdetail      0.0000000             0
## hvacheatingfuel        0.0000000             0
## construction           0.8638272     806426582
## plumbingfixturescount  0.0000000             0
## bathcount             15.4092955    6156214542
## bathpartialcount       0.0000000             0
## bedroomscount         14.2168193    6170513747
## roomscount             0.0000000             0
## storiescount          17.5233706    3054712470
## unitscount             7.0158528    2621327160
## fireplacecount         0.0000000             0
## roofmaterial          12.2232073    4885870364
## viewdescription        0.0000000             0
## porchcode              0.2536560     344471824
## porcharea              2.5403147    3033535350
## patioarea              5.0742273    2719591089
## deckflag               0.0000000             0
## deckarea               0.0000000             0
## drivewayarea           0.0000000             0
## pool                   4.1883868     772560000
## poolarea               8.9880421    2462694659
## fencearea              0.0000000             0
## arenaflag              0.0000000             0
## buildingscount         0.0000000             0
## shedcode               0.0000000             0
## utilitybuildingarea    0.0000000             0
## The test MSE is 122479255
```

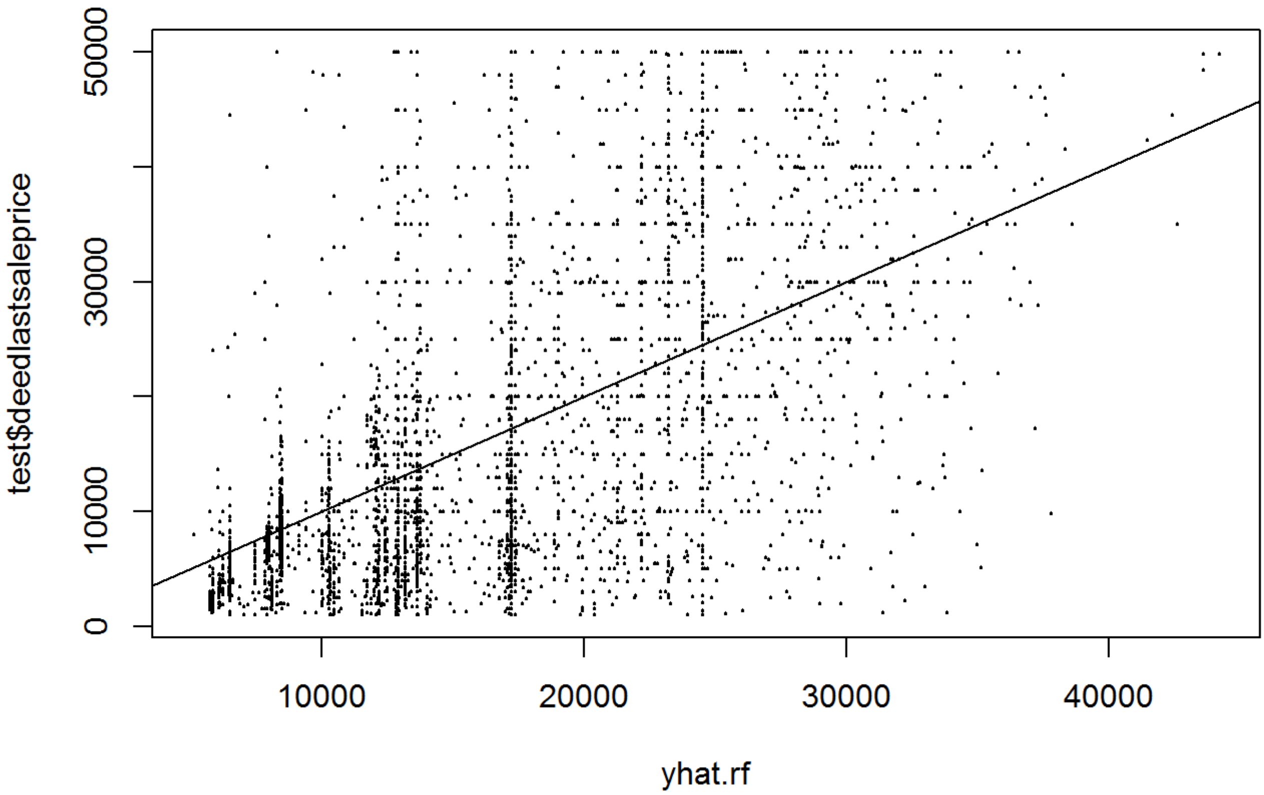


test$deedlastsaleprice
0
10000
30000
50000
10000
15000
20000
25000
30000
35000
40000
yhat.rf

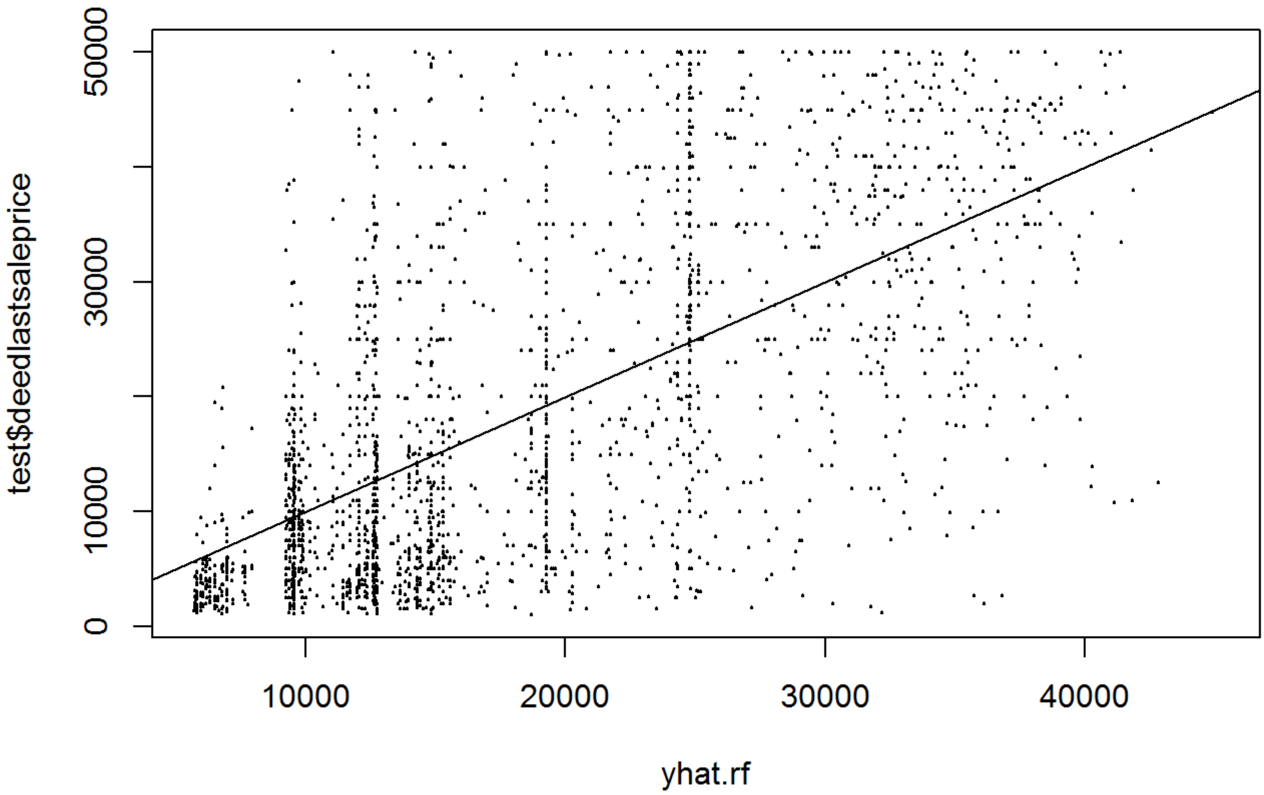
```
## NULL
## Below prints model for this state: FL and this county: Collier                         %IncMSE IncNo
dePurity
## ownertypedescription1 45.503263   89073798385
## ownertypedescription2 28.938286    8929389848
## parkinggarage          0.000000             0
## parkinggaragearea      0.000000             0
## hvacheatingdetail      0.000000             0
## hvacheatingfuel        0.000000             0
## plumbingfixturescount  0.000000             0
## bathcount              0.000000             0
## bathpartialcount       0.000000             0
## bedroomscount          0.000000             0
## roomscount             0.000000             0
## storiescount          27.765004   15807604265
## unitscount            43.062889   59213014492
## fireplacecount         0.000000             0
## roofmaterial           0.000000             0
## viewdescription        0.000000             0
## porchcode              3.684859     144292887
## porcharea             15.623643   10067614314
## patioarea              0.000000             0
## deckflag              10.145104    4434741268
## deckarea              13.545167    9433040689
## drivewayarea           0.000000             0
## pool                  12.419740     895430032
## poolarea               6.851176    3568552426
## fencearea              0.000000             0
## arenaflag              0.000000             0
## buildingscount         0.000000             0
## shedcode               0.000000             0
## utilitybuildingarea    0.000000             0
## The test MSE is 145528405
```


test$deedlastsaleprice
0
10000
30000
50000
10000
15000
20000
25000
30000
35000
40000
yhat.rf

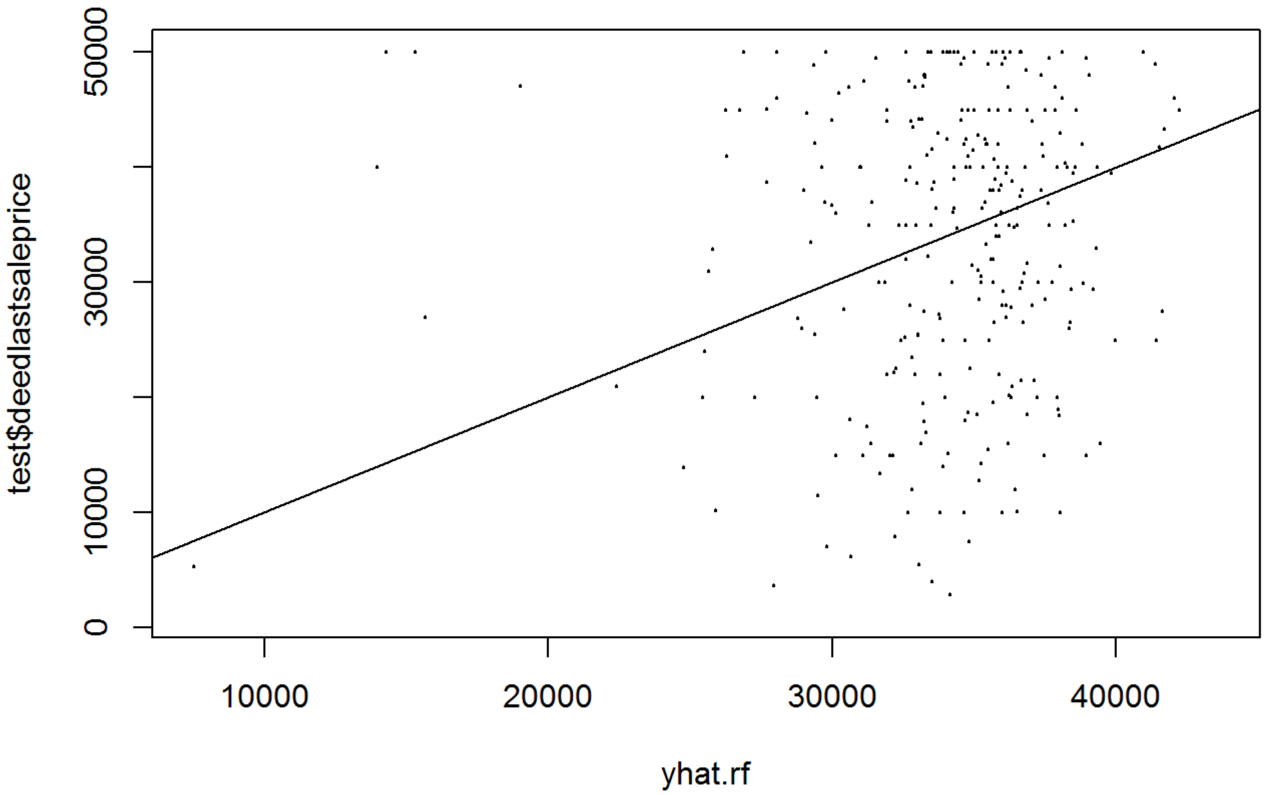
```
## NULL
## Below prints model for this state: FL and this county: Hendry                          %IncMSE IncNo
dePurity
## ownertypedescription1 31.6721561   17620163936
## ownertypedescription2 18.2988859   16885566594
## propertyusegroup      32.4822083   11933823920
## deedlastsaledate      68.0326326  427198067085
## parkinggarage          0.0000000             0
## parkinggaragearea      0.0000000             0
## hvacheatingdetail     14.7988666   17562242927
## hvacheatingfuel       12.9907014   23821184520
## construction          20.2218215   42092693219
## plumbingfixturescount  0.0000000             0
## bathcount             17.2890005   40953635945
## bathpartialcount       0.0000000             0
## bedroomscount         18.6027203   38367220168
## roomscount            -0.4661608    1617039565
## storiescount          26.9802212   62189302593
## unitscount            35.0912279   15251608864
## fireplacecount        -6.0125735    6261147501
## roofmaterial          17.7005110   54013308376
## viewdescription       -2.3291813      74057640
## porchcode              8.8921671    9845398139
## porcharea             15.7619146   78964442464
## patioarea              2.2487485   34013178451
## deckflag               0.0000000             0
## deckarea               0.0000000             0
## drivewayarea           0.0000000             0
## pool                  -2.6409595    1490511136
## poolarea              -5.0601824    2774110628
## fencearea              0.0000000             0
## arenaflag              0.0000000             0
## buildingscount         0.0000000             0
## shedcode               0.0000000             0
## utilitybuildingarea    0.0000000             0
## The test MSE is 103984787
```


test$deedlastsaleprice
0
10000
30000
50000
10000
20000
30000
40000
yhat.rf

```
## NULL
## Below prints model for this state: FL and this county: Lee                        %IncMSE IncNodePu
rity
## ownertypedescription1 27.493093    9314676455
## ownertypedescription2 20.622164   16111352256
## propertyusegroup      13.486437    3085075164
## deedlastsaledate      73.540096  193619101257
## parkinggaragearea     30.424879   35646162822
## hvacheatingdetail      9.585067    8694653484
## hvacheatingfuel       10.301742    9802573724
## plumbingfixturescount  0.000000             0
## bathcount             18.879314   85545388516
## bathpartialcount       0.000000             0
## bedroomscount         15.335332   58455695639
## roomscount             0.000000             0
## storiescount          22.619106   83450207838
## unitscount            18.054051   11949992051
## fireplacecount         3.399186    4149439002
## roofmaterial          10.002130   19346434500
## viewdescription       18.008248    7985879667
## porchcode              2.470582    8420039344
## porcharea             13.925962   71765075838
## patioarea             15.804661   21355564900
## deckflag               0.000000             0
## deckarea               0.000000             0
## drivewayarea           0.000000             0
## pool                  11.492480    3457934295
## poolarea              10.251958    7567060635
## fencearea              0.000000             0
## arenaflag              0.000000             0
## buildingscount         0.000000             0
## shedcode               0.000000             0
## utilitybuildingarea    0.000000             0
## The test MSE is 127366453
```

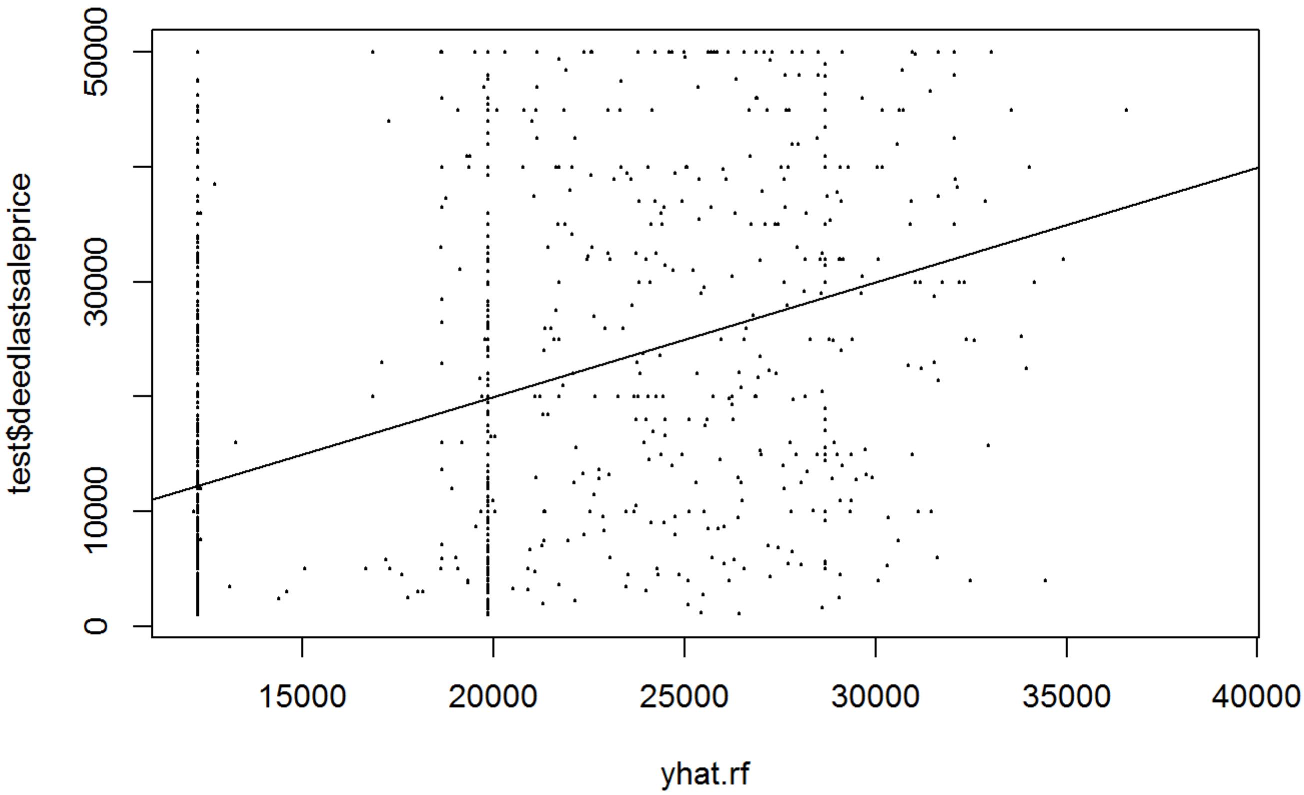

test$deedlastsaleprice
0
10000
30000
50000
10000
20000
30000
40000
yhat.rf

```
## NULL
## Below prints model for this state: FL and this county: Miami-Dade                    %IncMSE I
ncNodePurity
## ownertypedescription1  0.6094643    2442357696
## ownertypedescription2  4.6339532    3884121141
## yearbuilt             20.8161774   21786434919
## propertyusegroup       3.9548714    1008983236
## areabuilding          24.7980038   24751979036
## parkinggaragearea     -1.6273277    1402964954
## hvacheatingdetail     -2.8909768     173233894
## hvacheatingfuel        0.0000000             0
## plumbingfixturescount  0.0000000             0
## bathcount              4.8728255    3557784107
## bathpartialcount       0.0000000             0
## bedroomscount         15.2421450    5924683006
## roomscount             0.0000000             0
## storiescount          15.0609750    2547400488
## unitscount            10.3466887    1523547274
## fireplacecount         0.0000000             0
## roofmaterial           8.4987777    1091693207
## viewdescription        0.0000000             0
## porchcode              4.6617504     900515091
## porcharea             -0.0394279    3767201960
## patioarea              3.1045480    4644063539
## deckflag               0.0000000             0
## deckarea               0.0000000             0
## drivewayarea           0.0000000             0
## pool                   4.0756896     729767115
## poolarea              -1.6622019      61423260
## fencearea              0.0000000             0
## arenaflag              0.0000000             0
## buildingscount        -0.5669402     368729794
## shedcode               0.0000000             0
## utilitybuildingarea    0.0000000             0
## The test MSE is 162008275
```

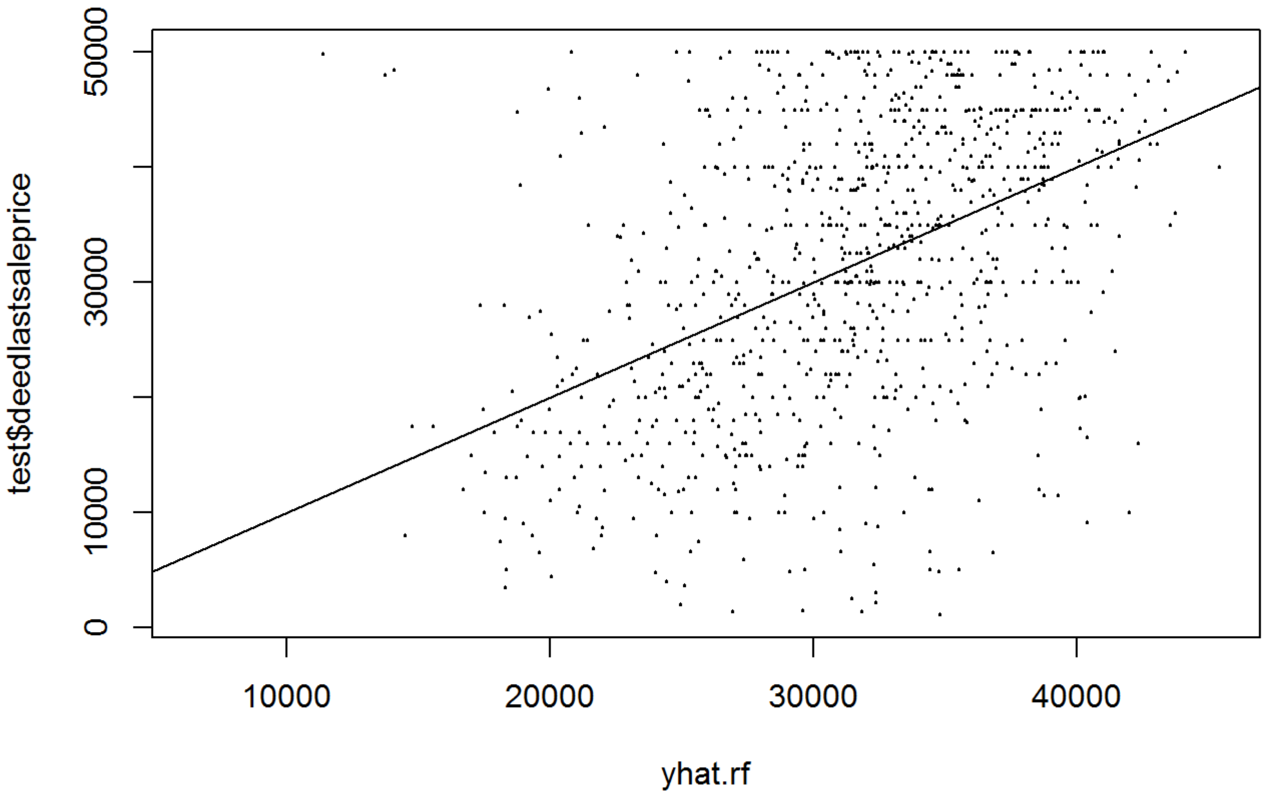


test$deedlastsaleprice
0
10000
30000
50000
10000
20000
30000
40000
yhat.rf

```
## NULL
## Below prints model for this state: FL and this county: Monroe                         %IncMSE IncNo
dePurity
## ownertypedescription1 56.1630807   49796600238
## ownertypedescription2 11.8778617    8080773499
## parkinggaragearea      0.7873837    7507430382
## hvacheatingdetail      0.3038532    3826198277
## hvacheatingfuel        0.0000000             0
## construction          -3.2079884     834448314
## plumbingfixturescount  0.0000000             0
## bathcount             10.0157583   11731555461
## bathpartialcount      -0.4381213    1465442905
## bedroomscount         13.1112088   22171172051
## roomscount             0.0000000             0
## storiescount          15.6373211   14932209630
## unitscount            36.5888934   27805098959
## fireplacecount         4.4096873     754240943
## roofmaterial           9.6502795   13296803551
## viewdescription        0.0000000             0
## porchcode              9.8030259    5007816355
## porcharea             17.3386922   40340970728
## patioarea             26.1879565   36598996852
## deckflag               8.0878657    3476602948
## deckarea               9.6570725   20132008435
## drivewayarea           0.0000000             0
## pool                  -0.2573755    1613536270
## poolarea              -3.9391708    2399465383
## fencearea              0.0000000             0
## arenaflag              0.0000000             0
## buildingscount         0.0000000             0
## shedcode               0.0000000             0
## utilitybuildingarea    0.0000000             0
## The test MSE is 194583573
```


test$deedlastsaleprice
0
10000
30000
50000
15000
20000
25000
30000
35000
40000
yhat.rf

```
## NULL
## Below prints model for this state: FL and this county: Palm Beach                         %IncMSE I
ncNodePurity
## ownertypedescription1  2.6014583    4925936451
## ownertypedescription2  5.7328689   11377591811
## yearbuilt             47.8258263   47068685028
## propertyusegroup       3.2110097     415108417
## deedlastsaledate      51.5885875   60601806422
## areabuilding          53.8466852   60511352727
## parkinggaragearea     18.1896562   13579686358
## hvacheatingdetail     13.4260520    3192846006
## hvacheatingfuel       11.5874129    3523122726
## construction          -2.4030600     870034708
## plumbingfixturescount  0.0000000             0
## bathcount             12.1964927    7428084550
## bathpartialcount      12.6510387    3722922913
## bedroomscount         24.9316273   13981426806
## roomscount            -1.0446488      43736003
## storiescount          12.0380663    3409185307
## unitscount             8.5612093    3701647645
## fireplacecount         0.0000000             0
## roofmaterial          16.4870565    9866732352
## viewdescription        0.1736852    2960366836
## porchcode             11.4849128    3641670723
## porcharea             34.9824658   30866454934
## patioarea             20.0691282   12977625045
## deckflag               5.8707607     679230604
## deckarea               8.3217129    1050140842
## drivewayarea           0.0000000             0
## pool                   6.4030731    1037805429
## poolarea               4.2517183    2696466081
## fencearea              0.0000000             0
## arenaflag              0.0000000             0
## buildingscount        11.3529687    2853381189
## shedcode               0.0000000             0
## utilitybuildingarea    0.0000000             0
## The test MSE is 123904377
```



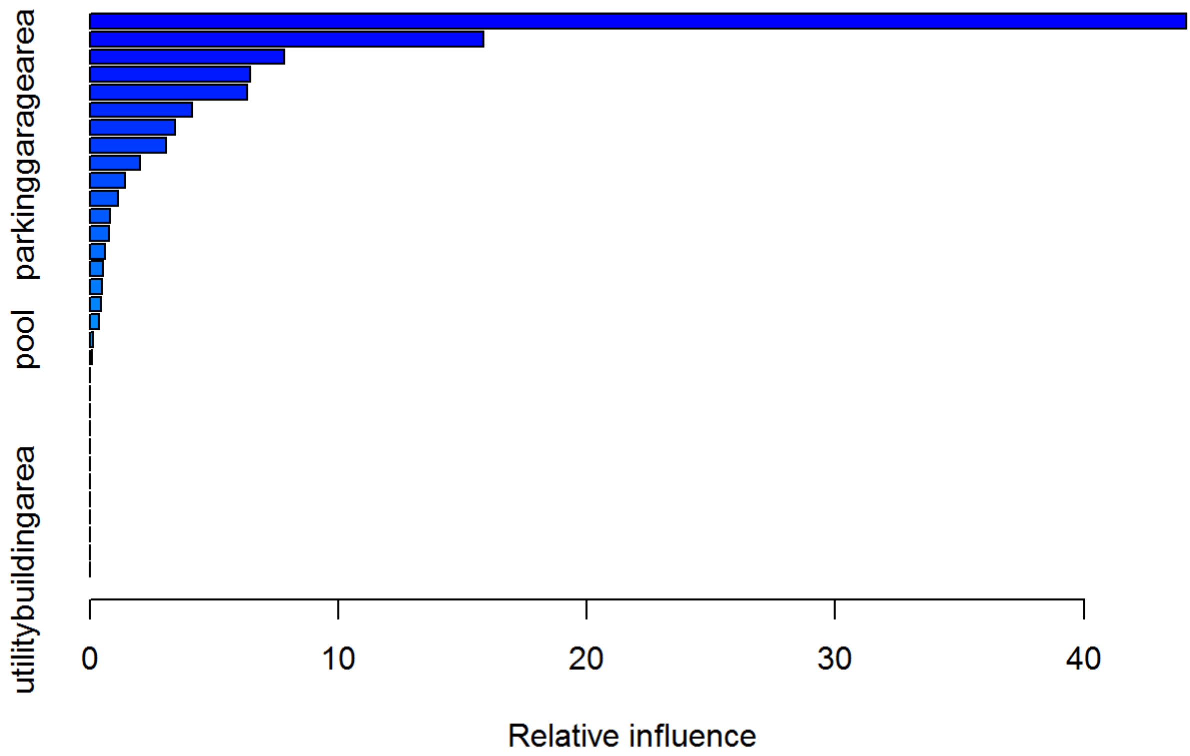

```
## NULL
```

```
## $Broward
## NULL
##
## $Collier
## NULL
##
## $Hendry
## NULL
##
## $Lee
## NULL
##
## $`Miami-Dade`
## NULL
##
## $Monroe
## NULL
##
## $`Palm Beach`
## NULL
```

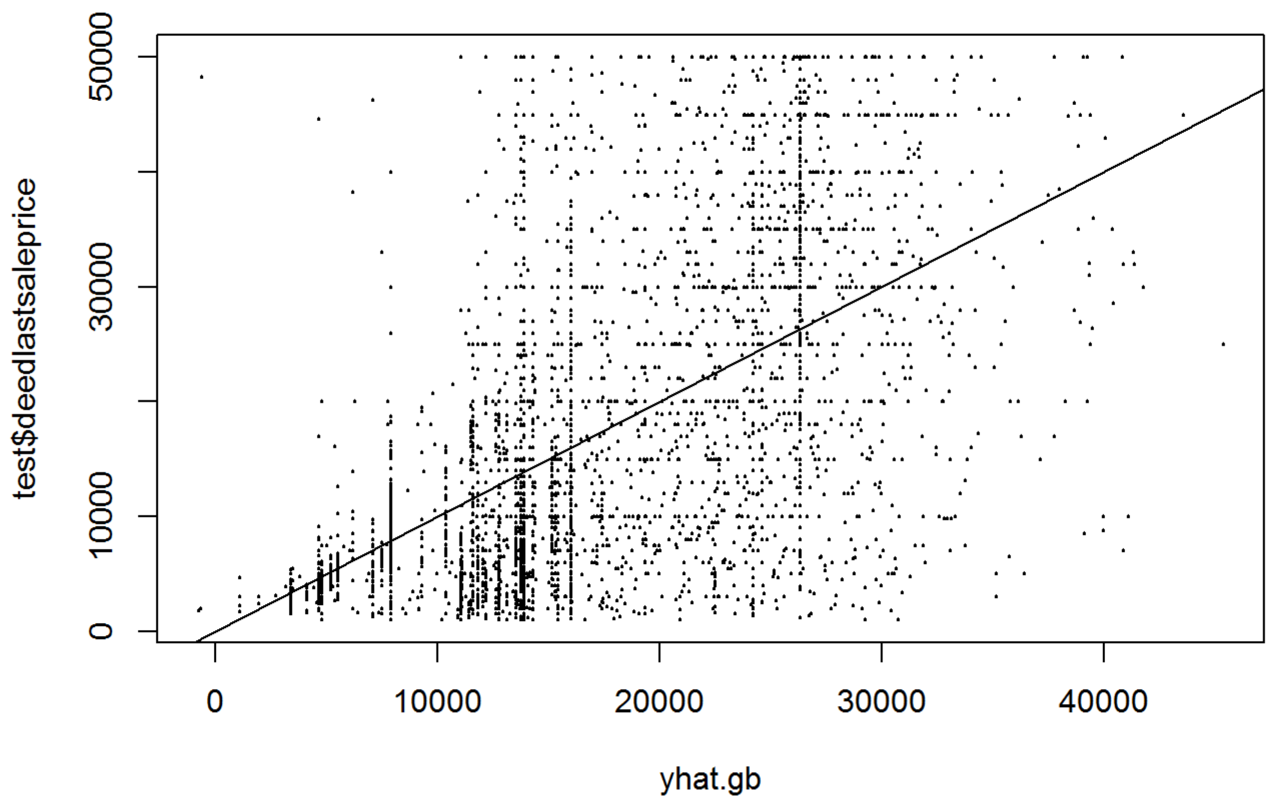
```
#Try boosting
gb.model<-function(dta){
#Report data subset
cat("Below prints model for this state:",
    dta$situsstatecode%>%unique()%>%as.character(),
    "and this county:",
    dta$situscounty%>%unique()%>%as.character()
)
#Reduce size (30000) for computation convenience
  if (nrow(dta)>30000){
    dta<-sample_n(dta,30000)
  }
#Excluding>$500000 transactions (extreme values) to experiment
dta$deedlastsaleprice[dta$deedlastsaleprice>50000]<-NA
#Excluding<$1000 transactions which are not authentic
dta$deedlastsaleprice[dta$deedlastsaleprice<1000]<-NA
dta$deedlastsaleprice<-dta$deedlastsaleprice%>%
  as.numeric()
dta<-filter(dta,is.na(dta$deedlastsaleprice)=="FALSE")
#Split training/test samples (0.7:0.3)
train<-sample_frac(dta,size=0.7)
test<-anti_join(dta,train,by="attomid")
#Missing value check (unused here)
check<-function(train){
for (i in 1:ncol(train)){
      a<-train[i]%>%nrow()
    b<-train[i]%>%is.na()%>%sum()
    c<-b/a
    print(c)
}
}
#Prepare y and x features
y<-train$deedlastsaleprice
x<-select(train,
          -attomid,
          -deedlastsaleprice,
          -situsstatecode,
          -situscounty)
#Gradient boosting
library(gbm)
train_gb <- gbm(y~.,x,
                n.trees = 1000,
                    distribution = "gaussian"
)
summary(train_gb)%>%print()
#Test set
yhat.gb <- predict(train_gb,test,n.trees = 1000)
cat("The test MSE is",
    mean((yhat.gb-test$deedlastsaleprice)^2,na.rm = TRUE)
    )
plot(yhat.gb,test$deedlastsaleprice,
     cex = .2)%>%print()
abline(0,1)
}
gb.model(home_county_cleaned_dta$Hendry)
```

```
## Below prints model for this state: FL and this county: Hendry
```

```
## Loaded gbm 2.1.5
```



```
##                                           var     rel.inf
## deedlastsaledate          deedlastsaledate 44.12180900
## porcharea                        porcharea 15.83869677
## storiescount                  storiescount  7.83794091
## hvacheatingfuel            hvacheatingfuel  6.43763148
## patioarea                        patioarea  6.34646034
## roofmaterial                  roofmaterial  4.11482934
## poolarea                          poolarea  3.43433132
## parkinggaragearea        parkinggaragearea  3.06831219
## propertyusegroup          propertyusegroup  2.02307901
## ownertypedescription1 ownertypedescription1  1.41714942
## construction                  construction  1.13251987
## fireplacecount              fireplacecount  0.81987415
## unitscount                      unitscount  0.77523336
## bathcount                        bathcount  0.61419827
## bedroomscount                bedroomscount  0.52629497
## hvacheatingdetail        hvacheatingdetail  0.47353029
## roomscount                      roomscount  0.43368080
## ownertypedescription2 ownertypedescription2  0.38521598
## pool                                  pool  0.12512875
## porchcode                        porchcode  0.07408379
## parkinggarage                parkinggarage  0.00000000
## plumbingfixturescount plumbingfixturescount  0.00000000
## bathpartialcount          bathpartialcount  0.00000000
## viewdescription            viewdescription  0.00000000
## deckflag                          deckflag  0.00000000
## deckarea                          deckarea  0.00000000
## drivewayarea                  drivewayarea  0.00000000
## fencearea                        fencearea  0.00000000
## arenaflag                        arenaflag  0.00000000
## buildingscount              buildingscount  0.00000000
## shedcode                          shedcode  0.00000000
## utilitybuildingarea    utilitybuildingarea  0.00000000
## The test MSE is 115601319
```




```
## NULL
```

```
#map(home_county_cleaned_dta, gb.model)
```

Transaction data

trx_dta<-fread("D:/raw_data/SF_Sales_Transactions_Data.csv")

trx_dta%>%group_by(attomid)%>%summarise(n())

The variable useful here is pretty straightforward: transaciton price 23 TransferAmount. After recoding it with "kick out <$1000" methods (this already kicks out nearly half in the sample) the result is still not so satisfying. Maybe think about adjust according to price per sq feet?

trx_dta$transferamount[trx_dta$transferamount<1000]<-NA trx_dta
$transferamount$transferamount%>%summary()

Regarding sales time, 2 variables look like compensating each other, but actually not sure: 13 InstrumentDate and 14 RecordingDate.

Parcel/risk data

risk_dta<-fread("D:/raw_data/SF_Parcel_Risk_and_Spatial_Data.csv")

One thing I'm still not clear is the geographical mapping that has been made. I understand one parcel division in the paper–properties with different levels of flooding risk. But you also mention Inverse Distance Weighted (IDW) to build value surfaces–is this automatic or manually tuned? Can we use it to construct neibourhood parcels that divides between models?

Besides, from 132 Totpopbg to 145 Hisptr seems to record ethnicity background information. How is that recorded and what’s the unit?

A series of interesting variables

for (i in 132:180){ a<-risk_dta[,i]%>%length() b<-risk_dta[,i]%>%is.na()%>%sum() if (b/a>0.1){ print(colnames(risk_dta[i])) } }

Apart from MedIncbg are all missing, other variables are pretty complete for analysis.