Final project for Math 390 Data Science at Queens College

Jianing Guo

5/18/2020

In collaboration with:

[Remessa] [Arnob] [Christella]

Abstraction:

All models are wrong, but some are useful. "Useful" is a such obscure word to be defined. Will the model come from Zillow(over $2.7 billions revenue company in year 2019) be "useful"? Obviously, if there is a better model than "zestimate", their model is not useful. Is there a better model? The following study will show you the answer.

Introduction:

Models are approximations or abstractions to the reality or phenomenon. To put it into simple language, model takes some important features (x's) from the data set as inputs to generate outputs (y's). If the error is low or under the controls, then it is useful model; vice versa. For this project, the mean goal is using algorithms to predict a future or new entry of an apartment selling price in Queens New York based on the raw data of Queens Apartments for sale from 2016 to 2017 scrapped from MLSLI using MTurk. The data has 55 features which are per house unit. The respond is the housing price. This project will be introducing three algorithms for prediction of the Sale Price through Linear Model, Regression Tree Model and Random Forest Model.

Data:

Our row data has 55 features, 2230 observations, and it includes 55 zip codes from Queens excluding Rockaways, a peninsula near JFK airport that is geographically distinct from the rest of the neighborhoods in Queens, NY. Some of those features are irrelevant to predict sale price. After removed those features, there are 25 features left. The following are the short summarize to those features: approx_year_built: integer; the apartment build recent years are more modern design and better equipment. If it is too old, it more likely to have high maintenance cost. community_district_num: integer; since some district has better school and facilities which will affect sale price. coop_condo: factor; coop have more community cost and owner only own the shares of stock from the corporation. On the other hand, condos are more expensive to own. The owner has the freedom to make changes and leasing. dining_room_type: factor; more room more expensive which usually mean larger in sq.ft. garage_exists: factor; which is a very important feature, due to the parking space is supper hard to find in NYC, especially, next to the apartment building. People who living in Flushing, Queens pay almost $200 per month for single park lot. kitchen_type: factor; here we factor kitchen into combo, eat in, efficiency, and none. This data has some spelling error we use sjmisc package to recode it. And we will to the same thing to garage_exists; factor it into 2 types: yes or no. num_bedrooms: integer, num_floors_in_building: integer, num_full_bathrooms: integer, num_half_bathrooms: integer, num_total_rooms: integer, parking_charges: factor, pct_tax_deductibl: integer, sq_footage, total_taxes: factor, walk_score: integer, monthly_cost: numeric, we combine maintenance_cost and common_charge. For address, we only extracted zip_code from the full address, since street name and building number don't relative to sale price. For car or dog allows, we simply combine into pet_alow, factor of yes and no. Those are the features which are

directly associate and correlated to the out put sale_price. The detail summary of those feature can be found in R Markdown line 92.

The missingness of this data set is quite interesting. There are 1702 NA's out of 2230 observations. Due to this large portion of missingness in the output columns, we have to split the data into real_data and fake_data. On the other hand, we use missing forest package to impute missing of the input features. We created m_s table with missing dummy variables for later comparison uses. In additional, we also created 80/20 split on the data for later verification.

Modeling:

Linear regression, we perform Ytrain based on Xtrain to get in-sample fit which has R^2 = 83% and Rmse = 94873.65. The coefficient for approx_year_build was -418, community_district_number was 1743, coop-condo was 192883, the price for square foot 435701.401, monthly_cost were 160, and parking_charges were 279 in unit of dollar. (for details see code line 174)

Regression Tree, Basic regression trees partition a data set into smaller groups and then fit a simple model (constant) for each subgroup. (UC Business Analytics R Programming Guide, n.d.) After tuning to get optimal tree model (tree with lowesr min_error). Our Roof node based on price_persqft which divided into two subgroups by price equal to 308000. And then, second layer nodes are sq_footage and num_total_room. (for more details see code line 259). For this model we have Rmse = 81942 and R^2 = 53.6%

Random Forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class

prediction and the class with the most votes becomes our model's prediction. (yiu, n.d.) For this model we have R^2 = 80.2% and Rmse = 76269.

Discussion

Due to significantly short on data, we only have 528 real data, to creating our model. Our model's R^2 are too low for production ready. In additional, the apartments near to LIRR or bus station will have relative more value than the apartments far away from it. I would like to research those stations and build into my models. I think this will boosting R^2 since we are equipped with more feature. To conclude, we are not going to beat Zillow this time.

Final project for Math 390 Data Science at Queens College

Jianing Guo

5/18/2020

In collaboration with: [Remessa] [Arnob] [Christella]

```
knitr::opts_chunk$set(echo = TRUE)
```

Load libraries

```
pkgs <- c('tidyverse', 'dplyr', 'tidyr', 'ggplot2', 'magrittr', 'stringr',
'mlr', 'sjmisc', 'missForest',
          'rsample', 'rpart', 'rpart.plot', 'ipred', 'caret')
for(p in pkgs) suppressPackageStartupMessages(stopifnot(
  library(p, quietly=TRUE,
          logical.return=TRUE,
          character.only=TRUE)))
```

Loading data

```
housing_data <- read.csv("housing_data_2016_2017.csv")
```

Remove features that will not be used and cleaning data

```
dat <- housing_data %>%
  select(-c(HITId, HITTypeId, Title, Description, Keywords, Reward,
CreationTime, MaxAssignments,   RequesterAnnotation,
AssignmentDurationInSeconds,    AutoApprovalDelayInSeconds, Expiration,
NumberOfSimilarHITs, LifetimeInSeconds, AssignmentId,   WorkerId,
AssignmentStatus,   AcceptTime, SubmitTime, AutoApprovalTime,   ApprovalTime,
RejectionTime,  RequesterFeedback,  WorkTimeInSeconds, LifetimeApprovalRate,
Last30DaysApprovalRate, Last7DaysApprovalRate, URL, url, date_of_sale))

dat.2 <- dat %>%
  mutate(zip_code = str_extract(full_address_or_zip_code, "[0-9]{5}"),
#extract 5-digit zipcode
         pets_allowed = ifelse((substr(cats_allowed, 1, 3) ==
"yes")|(substr(dogs_allowed, 1, 3) == "yes"), 1, 0),
          ) %>%
  select(-c(dogs_allowed,cats_allowed))#delete unwanted variables
```

convert currency columns into numeric for later calculation

```
dat.2$maintenance_cost <- as.numeric(gsub('\\$|,', '',
dat.2$maintenance_cost))
dat.2$common_charges <- as.numeric(gsub('\\$|,', '', dat.2$common_charges))
dat.2$parking_charges <- as.numeric(gsub('\\$|,', '', dat.2$parking_charges))
```

```
dat.2$listing_price_to_nearest_1000 <- as.numeric(gsub('\\$|,', '',
dat.2$listing_price_to_nearest_1000))
dat.2$total_taxes <- as.numeric(gsub('\\$|,', '', dat.2$total_taxes))
dat.2$sale_price <- as.numeric(gsub('\\$|,', '', dat.2$sale_price))
```

Use functions from sjmisc package to recode variables

```
dat.3 <- dat.2 %>%
  # recode all NAs with 0, and keep all others the same
  mutate(maintenance_cost = rec(maintenance_cost, rec = "NA = 0 ; else =
copy"),
         common_charges = rec(common_charges, rec = "NA = 0 ; else = copy"),
         # create a new variable monthly_cost to combine both costs
         monthly_cost = common_charges + maintenance_cost,
         # recode monthly_cost in case there's any NA
         monthly_cost = rec(monthly_cost, rec = "0 = NA ; else = copy"),
         # use rec function to recode garage exists variable
         garage_exists = rec(garage_exists, rec = "NA = 0 ; else = copy"),
         # garage exists variable needs further cleaning
         garage_exists = rec(garage_exists, rec = "NA = 0; eys , UG ,
Underground , yes , Yes = 1 ; else = copy"),
         garage_exists = as.factor(garage_exists),
         #use rec function to recode kitchen type
         kitchen_type = rec(kitchen_type, rec = "NA, none, 1955 = 0; combo,
Combo = 1; eat in, Eat in,  Eat In, eatin = 2; efficiemcy, efficiency,
efficiency kitchen, efficiency kitchene, efficiency ktchen = 3; else =
copy"),
         kitchen_type = as.factor(kitchen_type),
         # take care of some variable types
         dining_room_type = as.factor(dining_room_type),
         price_persqft = listing_price_to_nearest_1000 *1.0/ sq_footage,  #
sq_footage has NA??
         coop_condo = as.factor(tolower(coop_condo)),
         total_taxes = ifelse(total_taxes < 1000, NA, total_taxes)
         ) %>%
  # remove features that will not be used
  select(-c(maintenance_cost , common_charges, model_type, fuel_type,
zip_code, full_address_or_zip_code, listing_price_to_nearest_1000))
```

Do more data cleaning

```
dat.3 %<>%
  # create a ID column
  mutate(id = 1 : nrow(dat.3)) %>%
  # move id column to the first
  select(id, everything())
summary(dat.3)
```

```
##        id         approx_year_built community_district_num coop_condo
##  Min.   :   1.0   Min.   :1893      Min.   : 3.00          co-op:1661
##  1st Qu.: 558.2   1st Qu.:1950      1st Qu.:25.00          condo: 569
```

```
##  Median :1115.5   Median :1958      Median :26.00
##  Mean   :1115.5   Mean   :1963      Mean   :26.33
##  3rd Qu.:1672.8   3rd Qu.:1970      3rd Qu.:28.00
##  Max.   :2230.0   Max.   :2017      Max.   :32.00
##                   NA's   :40        NA's   :19
##     dining_room_type garage_exists kitchen_type  num_bedrooms
##  combo      :957     0:1826        0: 40        Min.   :0.000
##  dining area:  2     1: 404        1:399        1st Qu.:1.000
##  formal     :620                   2:942        Median :2.000
##  none       :  2                   3:849        Mean   :1.653
##  other      :201                                3rd Qu.:2.000
##  NA's       :448                                Max.   :6.000
##                                                 NA's   :115
##  num_floors_in_building num_full_bathrooms num_half_bathrooms
num_total_rooms
##  Min.   : 1.000         Min.   :1.000      Min.   :0.0000     Min.   :
0.000
##  1st Qu.: 3.000         1st Qu.:1.000      1st Qu.:1.0000     1st Qu.:
3.000
##  Median : 6.000         Median :1.000      Median :1.0000     Median :
4.000
##  Mean   : 7.785         Mean   :1.231      Mean   :0.9535     Mean   :
4.139
##  3rd Qu.: 7.000         3rd Qu.:1.000      3rd Qu.:1.0000     3rd Qu.:
5.000
##  Max.   :34.000         Max.   :3.000      Max.   :2.0000     Max.
:14.000
##  NA's   :650                               NA's   :2058       NA's   :2
##  parking_charges pct_tax_deductibl   sale_price       sq_footage
##  Min.   :  6.0   Min.   :20.0      Min.   : 55000   Min.   : 100.0
##  1st Qu.: 60.0   1st Qu.:40.0      1st Qu.:171500   1st Qu.: 743.0
##  Median : 99.0   Median :50.0      Median :259500   Median : 881.0
##  Mean   :107.6   Mean   :45.4      Mean   :314957   Mean   : 955.4
##  3rd Qu.:149.0   3rd Qu.:50.0      3rd Qu.:428875   3rd Qu.:1100.0
##  Max.   :837.0   Max.   :75.0      Max.   :999999   Max.   :6215.0
##  NA's   :1671    NA's   :1754      NA's   :1702     NA's   :1210
##   total_taxes     walk_score     pets_allowed     monthly_cost
##  Min.   :1024   Min.   : 7.00   Min.   :0.0000   Min.   : 100.0
##  1st Qu.:2500   1st Qu.:77.00   1st Qu.:0.0000   1st Qu.: 512.2
##  Median :3280   Median :89.00   Median :0.0000   Median : 687.0
##  Mean   :3412   Mean   :83.92   Mean   :0.3758   Mean   : 761.2
##  3rd Qu.:4104   3rd Qu.:95.00   3rd Qu.:1.0000   3rd Qu.: 911.5
##  Max.   :9300   Max.   :99.00   Max.   :1.0000   Max.   :4659.0
##  NA's   :1866                                    NA's   :100
##  price_persqft
##  Min.   :0.0579
##  1st Qu.:0.3223
##  Median :0.4146
##  Mean   :0.4675
##  3rd Qu.:0.5662
```

```
##  Max.   :3.1000
##  NA's   :1425

str(dat.3)

## 'data.frame':    2230 obs. of  21 variables:
##  $ id                    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ approx_year_built     : int  1955 1955 2004 2002 1949 1938 1950 1960
1960 2005 ...
##  $ community_district_num: int  25 25 24 25 26 28 29 28 25 30 ...
##  $ coop_condo            : Factor w/ 2 levels "co-op","condo": 1 1 2 2 1 1
1 1 1 2 ...
##  $ dining_room_type      : Factor w/ 5 levels "combo","dining area",..: 1
3 1 1 1 1 1 NA NA 5 ...
##  $ garage_exists         : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1
...
##  $ kitchen_type          : Factor w/ 4 levels "0","1","2","3": 3 3 4 3 3 3
4 4 3 3 ...
##  $ num_bedrooms          : int  2 1 1 3 2 2 1 0 1 1 ...
##  $ num_floors_in_building: int  6 7 1 NA 2 6 NA 2 NA 4 ...
##  $ num_full_bathrooms    : int  1 1 1 2 1 1 1 1 1 1 ...
##  $ num_half_bathrooms    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ num_total_rooms       : int  5 4 3 5 4 4 3 2 4 3 ...
##  $ parking_charges       : num  NA NA NA NA NA NA NA NA 20 NA ...
##  $ pct_tax_deductibl     : int  NA NA NA NA 39 NA NA NA NA NA ...
##  $ sale_price            : num  228000 235500 137550 545000 241700 ...
##  $ sq_footage            : int  NA 890 550 NA 675 1000 NA 375 NA 681 ...
##  $ total_taxes           : num  NA NA 5500 2260 NA NA NA NA NA 1320 ...
##  $ walk_score            : int  82 89 90 94 71 90 72 93 70 98 ...
##  $ pets_allowed          : num  0 0 0 0 1 1 0 0 0 0 ...
##  $ monthly_cost          : num  767 604 167 275 660 932 660 514 781 NA ...
##  $ price_persqft         : num  NA NA NA NA NA NA NA NA NA NA ...
```

Construct tables for modeling

```
real_y <- data.frame(dat.3$id, dat.3$sale_price)
real_dat <- subset(dat.3, (!is.na(dat.3$sale_price)))
fake_dat <- subset(dat.3, (is.na(dat.3$sale_price)))
real_dat$sale_price <- NULL
fake_dat$sale_price <- NULL
```

80/20 split on training and testing

```
train_indices <- sample(1 : nrow(real_dat), nrow(real_dat)*0.8)
training_data <- real_dat[train_indices, ]
testing_data <- real_dat[-train_indices, ]
X <- rbind(training_data, testing_data, fake_dat)
```

Create a table to store columns with missing data

```
m_d <- tbl_df(apply(is.na(X), 2, as.numeric))
colnames(m_d) <- paste("is_missing_", colnames(X), sep = "")
# remove duplicated rows
m_d <- tbl_df(t(unique(t(m_d))))
# remove rows where there is no missing data
m_d %<>% select_if(function(x){sum(x) > 0})
```

Data imputation

```
Ximp <- missForest(data.frame(X), sampsize = rep(172, ncol(X)))$ximp

##   missForest iteration 1 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want to
## do regression?

## done!
##   missForest iteration 2 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want to
## do regression?

## done!
##   missForest iteration 3 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want to
## do regression?

## done!
##   missForest iteration 4 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want to
## do regression?

## done!
##   missForest iteration 5 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want to
## do regression?
```

```
## done!
##   missForest iteration 6 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want to
## do regression?

## done!
##   missForest iteration 7 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want to
## do regression?

## done!
##   missForest iteration 8 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want to
## do regression?

## done!
##   missForest iteration 9 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want to
## do regression?

## done!

Ximp %<>%
  arrange(id)
```

Table with imputed data filled in

```
Xnew <- data.frame(cbind(Ximp, m_d, real_y))
Xnew %<>%
  mutate(price = dat.3.sale_price) %>%
  select(-c(id, dat.3.id, dat.3.sale_price))

linear_mod_impute_and_missing_dummies <- lm(price ~ ., data = Xnew)
summary(linear_mod_impute_and_missing_dummies)

##
## Call:
## lm(formula = price ~ ., data = Xnew)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -307097  -41488    2373   43477  301512
##
## Coefficients: (3 not defined because of singularities)
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.091e+06  5.861e+05   1.861  0.06327 .
## approx_year_built              -7.298e+02  3.009e+02  -2.425  0.01565 *
## community_district_num          2.137e+03  1.266e+03   1.689  0.09194 .
## coop_condocondo                 1.886e+05  1.860e+04  10.137  < 2e-16
***
## dining_room_typedining area     9.345e+03  5.838e+04   0.160  0.87288
## dining_room_typeformal          2.047e+04  9.442e+03   2.168  0.03060 *
## dining_room_typeother           2.890e+04  1.264e+04   2.286  0.02268 *
## garage_exists1                 -2.841e+03  9.887e+03  -0.287  0.77394
## kitchen_type1                   8.812e+03  3.227e+04   0.273  0.78493
## kitchen_type2                  -1.099e+03  3.146e+04  -0.035  0.97215
## kitchen_type3                  -1.444e+04  3.165e+04  -0.456  0.64827
## num_bedrooms                    4.970e+04  8.801e+03   5.647 2.76e-08
***
## num_floors_in_building          3.603e+03  7.830e+02   4.602 5.33e-06
***
## num_full_bathrooms              7.466e+04  1.300e+04   5.744 1.62e-08
***
## num_half_bathrooms              2.014e+04  2.464e+04   0.817  0.41404
## num_total_rooms                 2.355e+03  5.751e+03   0.409  0.68238
## parking_charges                 2.696e+02  1.083e+02   2.491  0.01308 *
## pct_tax_deductibl              -1.116e+03  1.234e+03  -0.905  0.36594
## sq_footage                      2.300e+01  1.420e+01   1.620  0.10597
## total_taxes                     1.454e+01  7.551e+00   1.926  0.05471 .
## walk_score                     -1.183e+02  3.335e+02  -0.355  0.72305
## pets_allowed                    2.178e+04  7.585e+03   2.872  0.00426
**
## monthly_cost                    1.287e+02  1.588e+01   8.104 4.20e-15
***
## price_persqft                   4.238e+05  7.366e+04   5.753 1.54e-08
***
## is_missing_approx_year_built    9.485e+03  3.418e+04   0.277  0.78153
## is_missing_community_district_num  2.045e+04  8.169e+04   0.250  0.80247
## is_missing_dining_room_type     1.461e+04  8.755e+03   1.669  0.09574 .
## is_missing_num_bedrooms                NA         NA      NA       NA
## is_missing_num_floors_in_building -1.436e+04  9.136e+03  -1.572  0.11655
## is_missing_num_half_bathrooms  -3.606e+04  1.570e+04  -2.296  0.02207 *
## is_missing_num_total_rooms             NA         NA      NA       NA
## is_missing_parking_charges     -8.625e+02  8.498e+03  -0.101  0.91920
## is_missing_pct_tax_deductibl   -2.622e+03  9.539e+03  -0.275  0.78353
## is_missing_sq_footage           1.523e+04  7.426e+03   2.051  0.04083 *
## is_missing_total_taxes          2.859e+03  1.012e+04   0.283  0.77759
## is_missing_monthly_cost         2.021e+04  2.192e+04   0.922  0.35688
## is_missing_price_persqft               NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 80360 on 494 degrees of freedom
##   (1702 observations deleted due to missingness)
## Multiple R-squared:  0.8122, Adjusted R-squared:  0.7996
## F-statistic: 64.73 on 33 and 494 DF,  p-value: < 2.2e-16
```

Take care of missing y

```
Data <- Xnew
# use imputed y to fill in sales price
Y <- Data$price
Data %<>%
  filter(!is.na(price)) %>%
  select(-price)
# 422: length of 80
Xtrain <- Data[1:422, ]
# real data row = 528
Xtest <- Data[423:528, ]
Ytrain <- Y[1:422]
Ytest <- Y[423:528]
```

Combine x/y train and x/y test

```
dtrain <- cbind(Xtrain, Ytrain)
dtest <- cbind(Xtest, Ytest)
```

Remove colinear features

```
Xtrain %<>%
  select(-c(is_missing_num_total_rooms, is_missing_num_bedrooms,
is_missing_price_persqft))
```

Simple linear regression

```
linear <- lm(Ytrain ~ ., data = Xtrain)
summary(linear)
```

```
##
## Call:
## lm(formula = Ytrain ~ ., data = Xtrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -304568  -40287    5060   40305  293193
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       1.075e+06  6.229e+05   1.727 0.085038 .
## approx_year_built                -7.108e+02  3.164e+02  -2.247 0.025217 *
## community_district_num            1.625e+03  1.316e+03   1.234 0.217797
## coop_condocondo                   2.031e+05  2.094e+04   9.698  < 2e-16
***
## dining_room_typedining area       1.653e+04  5.643e+04   0.293 0.769722
```

```
## dining_room_typeformal              2.224e+04  1.025e+04   2.169 0.030665 *
## dining_room_typeother               2.600e+04  1.345e+04   1.933 0.053907 .
## garage_exists1                      3.058e+01  1.101e+04   0.003 0.997785
## kitchen_type1                       3.117e+02  3.143e+04   0.010 0.992094
## kitchen_type2                      -4.172e+01  3.037e+04  -0.001 0.998905
## kitchen_type3                      -2.059e+04  3.065e+04  -0.672 0.502170
## num_bedrooms                        4.009e+04  9.551e+03   4.197 3.36e-05
***
## num_floors_in_building              3.128e+03  8.553e+02   3.657 0.000291
***
## num_full_bathrooms                  6.626e+04  1.408e+04   4.704 3.55e-06
***
## num_half_bathrooms                  1.115e+04  2.581e+04   0.432 0.665944
## num_total_rooms                     5.551e+03  6.290e+03   0.883 0.377982
## parking_charges                     3.059e+02  1.123e+02   2.723 0.006766
**
## pct_tax_deductibl                  -8.754e+02  1.527e+03  -0.573 0.566748
## sq_footage                          1.918e+01  1.467e+01   1.307 0.191962
## total_taxes                         1.281e+01  8.587e+00   1.492 0.136463
## walk_score                         -1.715e+02  3.527e+02  -0.486 0.627121
## pets_allowed                        1.320e+04  8.134e+03   1.623 0.105331
## monthly_cost                        1.575e+02  2.010e+01   7.834 4.58e-14
***
## price_persqft                       4.087e+05  8.185e+04   4.994 8.97e-07
***
## is_missing_approx_year_built        5.078e+03  4.084e+04   0.124 0.901109
## is_missing_community_district_num   2.054e+04  7.858e+04   0.261 0.793986
## is_missing_dining_room_type         1.305e+04  9.749e+03   1.338 0.181665
## is_missing_num_floors_in_building  -8.580e+03  9.953e+03  -0.862 0.389207
## is_missing_num_half_bathrooms      -3.125e+04  1.902e+04  -1.643 0.101252
## is_missing_parking_charges         -1.077e+04  8.985e+03  -1.199 0.231260
## is_missing_pct_tax_deductibl        1.169e+03  1.049e+04   0.111 0.911350
## is_missing_sq_footage               2.289e+04  8.020e+03   2.854 0.004546
**
## is_missing_total_taxes             -3.499e+01  1.082e+04  -0.003 0.997420
## is_missing_monthly_cost             1.407e+04  2.360e+04   0.596 0.551347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76980 on 388 degrees of freedom
## Multiple R-squared:  0.8254, Adjusted R-squared:  0.8106
## F-statistic: 55.59 on 33 and 388 DF,  p-value: < 2.2e-16
```

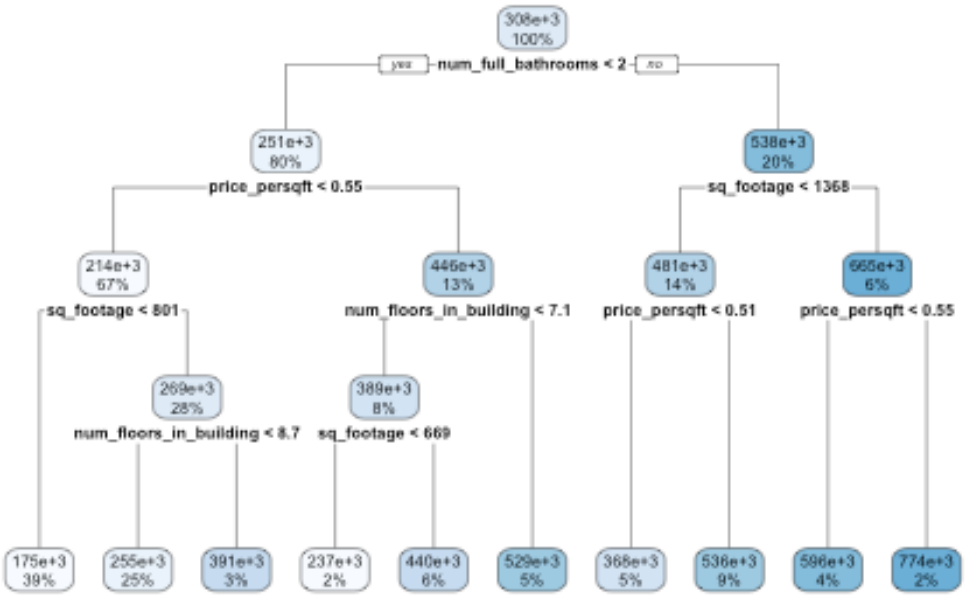
Make prediction， residuals， r^2

```
yhat <- predict(linear, Xtest)
e <- yhat - Ytest
sqrt(sum(e^2) / nrow(Xtest))
```

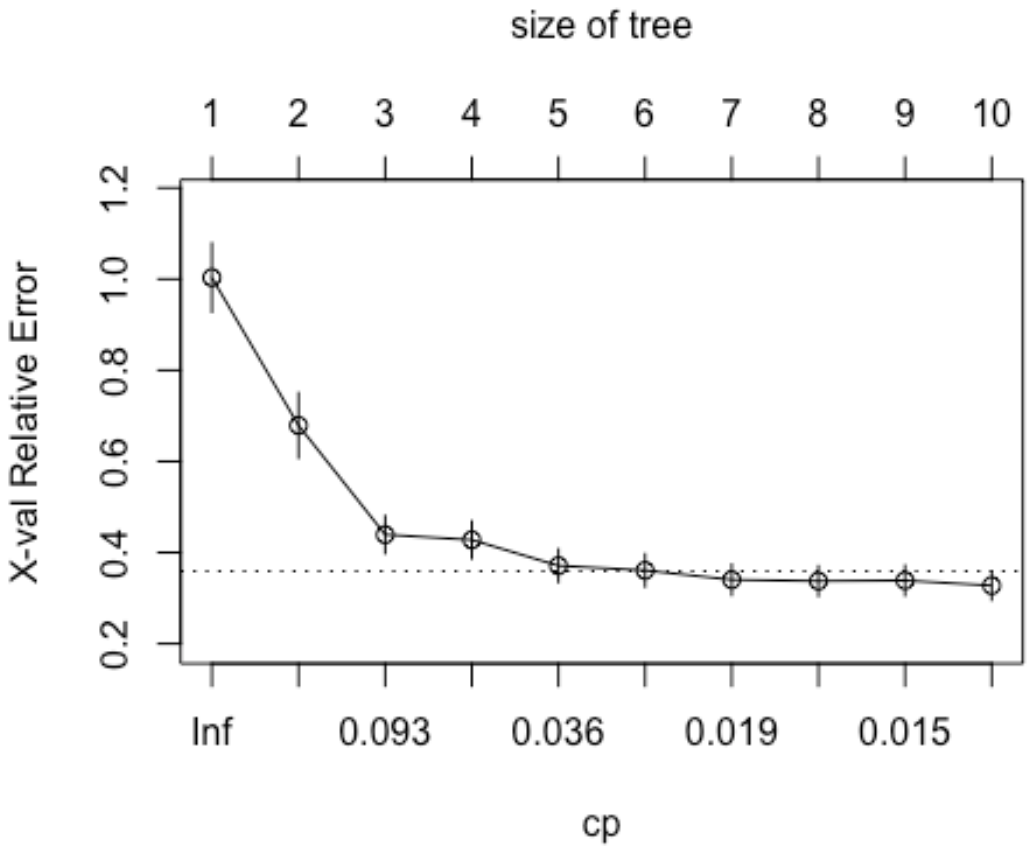
```
## [1] 95368.86
```

Regression tree mod 1

```
mod1 <- rpart(formula = Ytrain ~ .,
  data    = Xtrain,
  method  = "anova"
)
rpart.plot(mod1)
```




```
plotcp(mod1)
```




```
summary(mod1)
```

```
## Call:
## rpart(formula = Ytrain ~ ., data = Xtrain, method = "anova")
##   n= 422
##
##            CP nsplit rel error    xerror       xstd
## 1  0.42055632      0 1.0000000 1.0032124 0.07552372
## 2  0.18550984      1 0.5794437 0.6788940 0.07153101
## 3  0.04663919      2 0.3939338 0.4393409 0.04125317
## 4  0.04582955      3 0.3472947 0.4278045 0.04105597
## 5  0.02759198      4 0.3014651 0.3711104 0.03643080
## 6  0.01926389      5 0.2738731 0.3607922 0.03602838
## 7  0.01878043      6 0.2546092 0.3403475 0.03380166
## 8  0.01515856      7 0.2358288 0.3374488 0.03302046
## 9  0.01487063      8 0.2206703 0.3386975 0.03305337
## 10 0.01000000      9 0.2057996 0.3275400 0.03163880
##
## Variable importance
##     num_full_bathrooms             sq_footage            monthly_cost
##                     19                     17                       11
##          price_persqft            total_taxes         num_total_rooms
##                     10                      9                        8
```

```
##             coop_condo      approx_year_built     num_half_bathrooms
##                      7                      7                      4
##           num_bedrooms num_floors_in_building      pct_tax_deductibl
##                      3                      3                      2
##        parking_charges
##                      1
##
## Node number 1: 422 observations,    complexity param=0.4205563
##   mean=308191.7, MSE=3.121006e+10
##   left son=2 (338 obs) right son=3 (84 obs)
##   Primary splits:
##       num_full_bathrooms < 1.5       to the left,  improve=0.4205563, (0
missing)
##       price_persqft      < 0.5527845 to the left,  improve=0.3922844, (0
missing)
##       coop_condo         splits as  LR, improve=0.3754617, (0 missing)
##       approx_year_built  < 1970.5    to the left,  improve=0.3463094, (0
missing)
##       sq_footage         < 969.225   to the left,  improve=0.3139306, (0
missing)
##   Surrogate splits:
##       sq_footage         < 1133.495  to the left,  agree=0.917, adj=0.583,
(0 split)
##       total_taxes        < 4391.615  to the left,  agree=0.870, adj=0.345,
(0 split)
##       num_total_rooms    < 5.5       to the left,  agree=0.860, adj=0.298,
(0 split)
##       monthly_cost       < 1274.5    to the left,  agree=0.844, adj=0.214,
(0 split)
##       num_half_bathrooms < 0.455     to the right, agree=0.839, adj=0.190,
(0 split)
##
## Node number 2: 338 observations,    complexity param=0.1855098
##   mean=251077.9, MSE=1.663718e+10
##   left son=4 (284 obs) right son=5 (54 obs)
##   Primary splits:
##       price_persqft     < 0.5541679 to the left,  improve=0.4344879, (0
missing)
##       coop_condo        splits as  LR, improve=0.3876538, (0 missing)
##       approx_year_built < 1970.5    to the left,  improve=0.3067765, (0
missing)
##       sq_footage        < 857.955   to the left,  improve=0.1567627, (0
missing)
##       total_taxes       < 4198.69   to the left,  improve=0.1488824, (0
missing)
##   Surrogate splits:
##       coop_condo        splits as  LR, agree=0.956, adj=0.722, (0 split)
##       approx_year_built < 1970.5    to the left,  agree=0.944, adj=0.648,
(0 split)
##       monthly_cost      < 390.5     to the right, agree=0.902, adj=0.389,
```

```
(0 split)
##       pct_tax_deductibl < 34.945    to the right, agree=0.864, adj=0.148,
(0 split)
##       total_taxes       < 4463.51   to the left,  agree=0.861, adj=0.130,
(0 split)
##
## Node number 3: 84 observations,    complexity param=0.04582955
##   mean=538006.5, MSE=2.390811e+10
##   left son=6 (58 obs) right son=7 (26 obs)
##   Primary splits:
##       sq_footage             < 1368.09   to the left,  improve=0.3005579,
(0 missing)
##       price_persqft          < 0.4700775 to the left,  improve=0.2564064,
(0 missing)
##       num_floors_in_building < 14.5      to the left,  improve=0.2206020,
(0 missing)
##       parking_charges        < 69.735    to the left,  improve=0.2094910,
(0 missing)
##       total_taxes            < 4976.01   to the left,  improve=0.2043087,
(0 missing)
##   Surrogate splits:
##       monthly_cost           < 1461.5    to the left,  agree=0.857,
adj=0.538, (0 split)
##       num_bedrooms           < 2.5       to the left,  agree=0.833,
adj=0.462, (0 split)
##       num_floors_in_building < 11.525    to the left,  agree=0.798,
adj=0.346, (0 split)
##       num_total_rooms        < 6.5       to the left,  agree=0.774,
adj=0.269, (0 split)
##       total_taxes            < 4787.5    to the left,  agree=0.762,
adj=0.231, (0 split)
##
## Node number 4: 284 observations,    complexity param=0.04663919
##   mean=214004.2, MSE=7.534685e+09
##   left son=8 (166 obs) right son=9 (118 obs)
##   Primary splits:
##       sq_footage      < 800.5     to the left,  improve=0.2870613, (0
missing)
##       monthly_cost    < 761.5     to the left,  improve=0.2363109, (0
missing)
##       num_bedrooms    < 1.5       to the left,  improve=0.2009774, (0
missing)
##       num_total_rooms < 4.5       to the left,  improve=0.1218869, (0
missing)
##       price_persqft   < 0.5121664 to the left,  improve=0.1065049, (0
missing)
##   Surrogate splits:
##       num_bedrooms     < 1.5       to the left,  agree=0.891, adj=0.737,
(0 split)
##       num_total_rooms  < 3.5       to the left,  agree=0.835, adj=0.602,
```

```
(0 split)
##       monthly_cost     < 761.5     to the left,  agree=0.803, adj=0.525,
(0 split)
##       total_taxes      < 2831.22   to the left,  agree=0.704, adj=0.288,
(0 split)
##       dining_room_type splits as  LLR-R, agree=0.648, adj=0.153, (0 split)
##
## Node number 5: 54 observations,    complexity param=0.01926389
##   mean=446058.3, MSE=1.926355e+10
##   left son=10 (32 obs) right son=11 (22 obs)
##   Primary splits:
##       num_floors_in_building < 7.125     to the left,  improve=0.2439052,
(0 missing)
##       monthly_cost           < 304       to the left,  improve=0.2437085,
(0 missing)
##       total_taxes            < 2503.925  to the left,  improve=0.2306287,
(0 missing)
##       num_half_bathrooms     < 1.015     to the right, improve=0.2085708,
(0 missing)
##       sq_footage             < 669       to the left,  improve=0.2063172,
(0 missing)
##   Surrogate splits:
##       monthly_cost       < 401.5     to the left,  agree=0.796, adj=0.500,
(0 split)
##       num_half_bathrooms < 1.005     to the right, agree=0.741, adj=0.364,
(0 split)
##       pct_tax_deductibl  < 47.865    to the right, agree=0.722, adj=0.318,
(0 split)
##       total_taxes        < 4255.025  to the left,  agree=0.722, adj=0.318,
(0 split)
##       parking_charges    < 135.86    to the right, agree=0.685, adj=0.227,
(0 split)
##
## Node number 6: 58 observations,    complexity param=0.02759198
##   mean=481250.9, MSE=1.432215e+10
##   left son=12 (19 obs) right son=13 (39 obs)
##   Primary splits:
##       price_persqft     < 0.5087043 to the left,  improve=0.4374757, (0
missing)
##       approx_year_built < 1966.5    to the left,  improve=0.4026342, (0
missing)
##       coop_condo        splits as  LR, improve=0.3440129, (0 missing)
##       monthly_cost      < 812       to the right, improve=0.3440129, (0
missing)
##       parking_charges   < 73.5725   to the left,  improve=0.3059630, (0
missing)
##   Surrogate splits:
##       coop_condo        splits as  LR, agree=0.948, adj=0.842, (0 split)
##       monthly_cost      < 812       to the right, agree=0.948, adj=0.842,
(0 split)
```

```
##       approx_year_built < 1963.5    to the left,  agree=0.914, adj=0.737,
(0 split)
##       sq_footage        < 1143.5    to the right, agree=0.862, adj=0.579,
(0 split)
##       parking_charges   < 73.5725   to the left,  agree=0.793, adj=0.368,
(0 split)
##
## Node number 7: 26 observations,    complexity param=0.01487063
##   mean=664615.4, MSE=2.207662e+10
##   left son=14 (16 obs) right son=15 (10 obs)
##   Primary splits:
##       price_persqft          < 0.5451458 to the left,  improve=0.3412168,
(0 missing)
##       total_taxes            < 5018.9    to the left,  improve=0.3039461,
(0 missing)
##       parking_charges        < 141.8825  to the left,  improve=0.2528833,
(0 missing)
##       community_district_num < 27.5      to the left,  improve=0.2358061,
(0 missing)
##       num_half_bathrooms     < 0.945     to the right, improve=0.1572672,
(0 missing)
##   Surrogate splits:
##       total_taxes            < 5018.9    to the left,  agree=0.808,
adj=0.5, (0 split)
##       num_floors_in_building < 21        to the left,  agree=0.769,
adj=0.4, (0 split)
##       approx_year_built      < 1970.5    to the left,  agree=0.731,
adj=0.3, (0 split)
##       pct_tax_deductibl      < 35.69     to the right, agree=0.731,
adj=0.3, (0 split)
##       monthly_cost           < 2211      to the left,  agree=0.731,
adj=0.3, (0 split)
##
## Node number 8: 166 observations
##   mean=174793.3, MSE=3.347287e+09
##
## Node number 9: 118 observations,    complexity param=0.01515856
##   mean=269165.3, MSE=8.219769e+09
##   left son=18 (106 obs) right son=19 (12 obs)
##   Primary splits:
##       num_floors_in_building < 8.705     to the left,  improve=0.2058369,
(0 missing)
##       parking_charges        < 88.07     to the left,  improve=0.1792023,
(0 missing)
##       price_persqft          < 0.4516464 to the left,  improve=0.1596936,
(0 missing)
##       walk_score             < 91.5      to the left,  improve=0.1334127,
(0 missing)
##       monthly_cost           < 1048      to the left,  improve=0.1099628,
(0 missing)
```

```
##   Surrogate splits:
##       parking_charges   < 395.855   to the left,  agree=0.915, adj=0.167,
(0 split)
##       approx_year_built < 1964.5    to the left,  agree=0.907, adj=0.083,
(0 split)
##
## Node number 10: 32 observations,    complexity param=0.01878043
##   mean=389223.4, MSE=1.593423e+10
##   left son=20 (8 obs) right son=21 (24 obs)
##   Primary splits:
##       sq_footage      < 669       to the left,  improve=0.4851004, (0
missing)
##       num_total_rooms < 3.5       to the left,  improve=0.2499396, (0
missing)
##       parking_charges < 143.74    to the right, improve=0.2106907, (0
missing)
##       monthly_cost    < 297       to the left,  improve=0.1928689, (0
missing)
##       total_taxes     < 2503.925  to the left,  improve=0.1837103, (0
missing)
##   Surrogate splits:
##       num_bedrooms    < 0.5       to the left,  agree=0.812, adj=0.250, (0
split)
##       monthly_cost    < 177       to the left,  agree=0.812, adj=0.250, (0
split)
##       num_total_rooms < 2.5       to the left,  agree=0.781, adj=0.125, (0
split)
##       price_persqft   < 0.7280547 to the right, agree=0.781, adj=0.125, (0
split)
##
## Node number 11: 22 observations
##   mean=528727.3, MSE=1.257356e+10
##
## Node number 12: 19 observations
##   mean=367844.7, MSE=8.995931e+09
##
## Node number 13: 39 observations
##   mean=536500, MSE=7.59891e+09
##
## Node number 14: 16 observations
##   mean=596000, MSE=1.695512e+10
##
## Node number 15: 10 observations
##   mean=774400, MSE=1.068544e+10
##
## Node number 18: 106 observations
##   mean=255325.5, MSE=6.709224e+09
##
## Node number 19: 12 observations
##   mean=391416.7, MSE=4.925576e+09
```

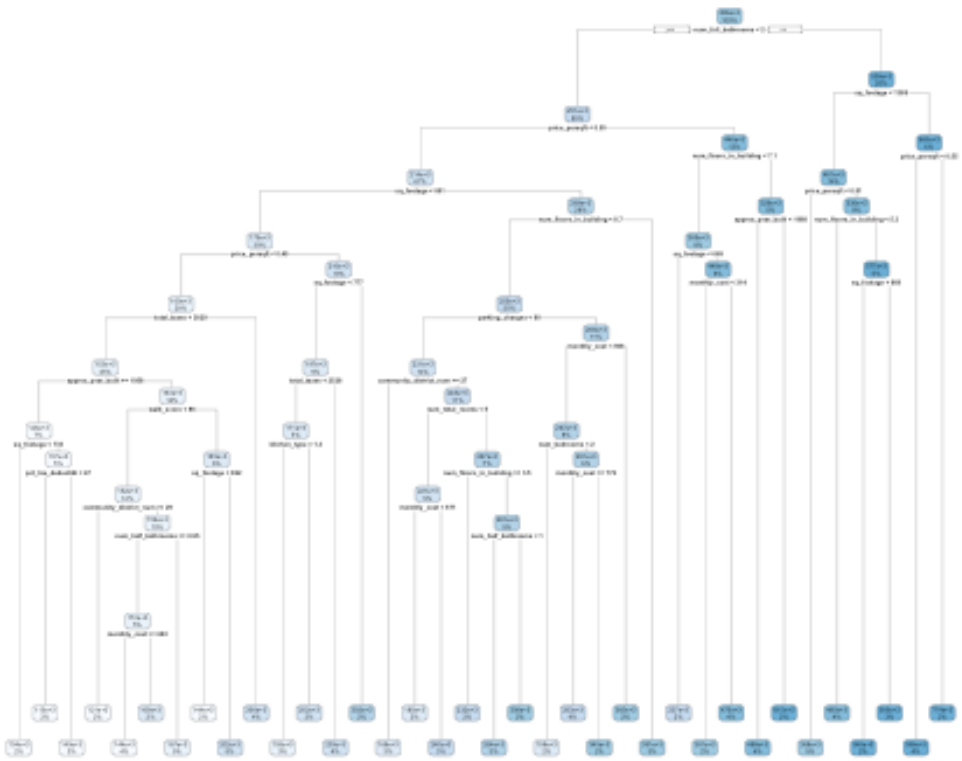
```
##
## Node number 20: 8 observations
##   mean=236943.8, MSE=5.856097e+09
##
## Node number 21: 24 observations
##   mean=439983.3, MSE=8.987341e+09
```

```
yhat <- predict(mod1, Xtest)
e <- yhat - Ytest
# 106: length of testing table
sqrt(sum(e^2)/106)
```

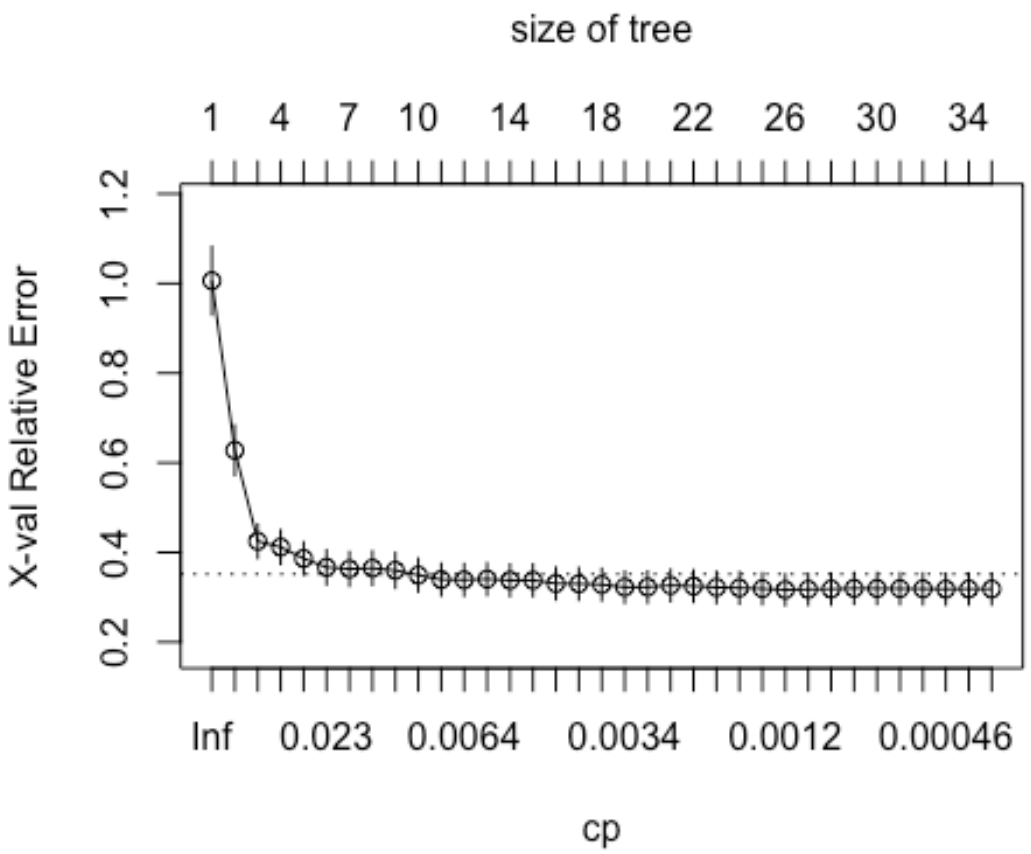
```
## [1] 100728.2
```

###mode 2

```
mod2 <- rpart(
  formula = Ytrain ~ .,
  data    = Xtrain,
  method  = "anova",
  control = list(cp = 0, xval = 10)
)
rpart.plot(mod2)
```



```r
plotcp(mod2)
```




```r
summary(mod2)
```

```
## Call:
## rpart(formula = Ytrain ~ ., data = Xtrain, method = "anova",
##     control = list(cp = 0, xval = 10))
##   n= 422
##
##              CP nsplit rel error    xerror       xstd
## 1  0.4205563162      0 1.0000000 1.0062184 0.07592906
## 2  0.1855098366      1 0.5794437 0.6278150 0.05586384
## 3  0.0466391851      2 0.3939338 0.4250392 0.03783441
## 4  0.0458295475      3 0.3472947 0.4118463 0.03815499
## 5  0.0275919822      4 0.3014651 0.3863977 0.03657124
## 6  0.0192638942      5 0.2738731 0.3667320 0.03880221
## 7  0.0187804315      6 0.2546092 0.3631764 0.03807725
## 8  0.0151585551      7 0.2358288 0.3648373 0.03806376
## 9  0.0148706251      8 0.2206703 0.3600266 0.03937456
## 10 0.0079022374      9 0.2057996 0.3493527 0.03731881
## 11 0.0066764275     10 0.1978974 0.3393717 0.03596327
## 12 0.0061801498     11 0.1912210 0.3386993 0.03600030
## 13 0.0058018889     12 0.1850408 0.3404516 0.03604850
```

```
## 14 0.0056075137     13 0.1792389 0.3377880 0.03592434
## 15 0.0054916570     14 0.1736314 0.3377880 0.03592434
## 16 0.0049499104     15 0.1681398 0.3307467 0.03604634
## 17 0.0044126359     16 0.1631898 0.3300093 0.03600777
## 18 0.0037737326     17 0.1587772 0.3282174 0.03594620
## 19 0.0030509570     18 0.1550035 0.3223701 0.03540815
## 20 0.0027279436     19 0.1519525 0.3226547 0.03540908
## 21 0.0023364086     20 0.1492246 0.3263320 0.03562167
## 22 0.0021828030     21 0.1468882 0.3248034 0.03563313
## 23 0.0018814318     22 0.1447054 0.3219394 0.03556260
## 24 0.0014245176     23 0.1428239 0.3207396 0.03565157
## 25 0.0013994727     24 0.1413994 0.3190699 0.03501429
## 26 0.0010570011     25 0.1399999 0.3166525 0.03503353
## 27 0.0010337534     26 0.1389429 0.3176712 0.03506962
## 28 0.0009341672     27 0.1379092 0.3182811 0.03505865
## 29 0.0007780139     28 0.1369750 0.3198469 0.03500553
## 30 0.0007767720     29 0.1361970 0.3195437 0.03500680
## 31 0.0006390631     30 0.1354202 0.3191796 0.03499311
## 32 0.0005192759     31 0.1347812 0.3187092 0.03501396
## 33 0.0004022024     32 0.1342619 0.3180782 0.03501471
## 34 0.0002183995     33 0.1338597 0.3184845 0.03500652
## 35 0.0000000000     35 0.1334229 0.3182978 0.03500927
##
## Variable importance
##     num_full_bathrooms             sq_footage           monthly_cost
##                     17                     16                     11
##          price_persqft            total_taxes        num_total_rooms
##                     11                      9                      7
##      approx_year_built             coop_condo     num_half_bathrooms
##                      7                      7                      4
## num_floors_in_building           num_bedrooms      pct_tax_deductibl
##                      3                      3                      2
##        parking_charges community_district_num
##                      1                      1
##
## Node number 1: 422 observations,    complexity param=0.4205563
##   mean=308191.7, MSE=3.121006e+10
##   left son=2 (338 obs) right son=3 (84 obs)
##   Primary splits:
##       num_full_bathrooms < 1.5       to the left,  improve=0.4205563, (0
missing)
##       price_persqft      < 0.5527845 to the left,  improve=0.3922844, (0
missing)
##       coop_condo         splits as  LR, improve=0.3754617, (0 missing)
##       approx_year_built  < 1970.5    to the left,  improve=0.3463094, (0
missing)
##       sq_footage         < 969.225   to the left,  improve=0.3139306, (0
missing)
##   Surrogate splits:
##       sq_footage         < 1133.495  to the left,  agree=0.917, adj=0.583,
```

```
(0 split)
##       total_taxes        < 4391.615  to the left,  agree=0.870, adj=0.345,
(0 split)
##       num_total_rooms    < 5.5       to the left,  agree=0.860, adj=0.298,
(0 split)
##       monthly_cost       < 1274.5    to the left,  agree=0.844, adj=0.214,
(0 split)
##       num_half_bathrooms < 0.455     to the right, agree=0.839, adj=0.190,
(0 split)
##
## Node number 2: 338 observations,    complexity param=0.1855098
##   mean=251077.9, MSE=1.663718e+10
##   left son=4 (284 obs) right son=5 (54 obs)
##   Primary splits:
##       price_persqft     < 0.5541679 to the left,  improve=0.4344879, (0
missing)
##       coop_condo        splits as  LR, improve=0.3876538, (0 missing)
##       approx_year_built < 1970.5    to the left,  improve=0.3067765, (0
missing)
##       sq_footage        < 857.955   to the left,  improve=0.1567627, (0
missing)
##       total_taxes       < 4198.69   to the left,  improve=0.1488824, (0
missing)
##   Surrogate splits:
##       coop_condo        splits as  LR, agree=0.956, adj=0.722, (0 split)
##       approx_year_built < 1970.5    to the left,  agree=0.944, adj=0.648,
(0 split)
##       monthly_cost      < 390.5     to the right, agree=0.902, adj=0.389,
(0 split)
##       pct_tax_deductibl < 34.945    to the right, agree=0.864, adj=0.148,
(0 split)
##       total_taxes       < 4463.51   to the left,  agree=0.861, adj=0.130,
(0 split)
##
## Node number 3: 84 observations,    complexity param=0.04582955
##   mean=538006.5, MSE=2.390811e+10
##   left son=6 (58 obs) right son=7 (26 obs)
##   Primary splits:
##       sq_footage            < 1368.09   to the left,  improve=0.3005579,
(0 missing)
##       price_persqft         < 0.4700775 to the left,  improve=0.2564064,
(0 missing)
##       num_floors_in_building < 14.5      to the left,  improve=0.2206020,
(0 missing)
##       parking_charges       < 69.735    to the left,  improve=0.2094910,
(0 missing)
##       total_taxes           < 4976.01   to the left,  improve=0.2043087,
(0 missing)
##   Surrogate splits:
##       monthly_cost          < 1461.5    to the left,  agree=0.857,
```

```
adj=0.538, (0 split)
##       num_bedrooms           < 2.5       to the left,  agree=0.833,
adj=0.462, (0 split)
##       num_floors_in_building < 11.525    to the left,  agree=0.798,
adj=0.346, (0 split)
##       num_total_rooms        < 6.5       to the left,  agree=0.774,
adj=0.269, (0 split)
##       total_taxes            < 4787.5    to the left,  agree=0.762,
adj=0.231, (0 split)
##
## Node number 4: 284 observations,    complexity param=0.04663919
##   mean=214004.2, MSE=7.534685e+09
##   left son=8 (166 obs) right son=9 (118 obs)
##   Primary splits:
##       sq_footage      < 800.5     to the left,  improve=0.2870613, (0
missing)
##       monthly_cost    < 761.5     to the left,  improve=0.2363109, (0
missing)
##       num_bedrooms    < 1.5       to the left,  improve=0.2009774, (0
missing)
##       num_total_rooms < 4.5       to the left,  improve=0.1218869, (0
missing)
##       price_persqft   < 0.5121664 to the left,  improve=0.1065049, (0
missing)
##   Surrogate splits:
##       num_bedrooms     < 1.5       to the left,  agree=0.891, adj=0.737,
(0 split)
##       num_total_rooms  < 3.5       to the left,  agree=0.835, adj=0.602,
(0 split)
##       monthly_cost     < 761.5     to the left,  agree=0.803, adj=0.525,
(0 split)
##       total_taxes      < 2831.22   to the left,  agree=0.704, adj=0.288,
(0 split)
##       dining_room_type splits as  LLR-R, agree=0.648, adj=0.153, (0 split)
##
## Node number 5: 54 observations,    complexity param=0.01926389
##   mean=446058.3, MSE=1.926355e+10
##   left son=10 (32 obs) right son=11 (22 obs)
##   Primary splits:
##       num_floors_in_building < 7.125     to the left,  improve=0.2439052,
(0 missing)
##       monthly_cost           < 304       to the left,  improve=0.2437085,
(0 missing)
##       total_taxes            < 2503.925  to the left,  improve=0.2306287,
(0 missing)
##       num_half_bathrooms     < 1.015     to the right, improve=0.2085708,
(0 missing)
##       sq_footage             < 669       to the left,  improve=0.2063172,
(0 missing)
##   Surrogate splits:
```

```
##       monthly_cost       < 401.5     to the left,  agree=0.796, adj=0.500,
(0 split)
##       num_half_bathrooms < 1.005     to the right, agree=0.741, adj=0.364,
(0 split)
##       pct_tax_deductibl  < 47.865    to the right, agree=0.722, adj=0.318,
(0 split)
##       total_taxes        < 4255.025  to the left,  agree=0.722, adj=0.318,
(0 split)
##       parking_charges    < 135.86    to the right, agree=0.685, adj=0.227,
(0 split)
##
## Node number 6: 58 observations,    complexity param=0.02759198
##   mean=481250.9, MSE=1.432215e+10
##   left son=12 (19 obs) right son=13 (39 obs)
##   Primary splits:
##       price_persqft     < 0.5087043 to the left,  improve=0.4374757, (0
missing)
##       approx_year_built < 1966.5    to the left,  improve=0.4026342, (0
missing)
##       coop_condo        splits as  LR, improve=0.3440129, (0 missing)
##       monthly_cost      < 812       to the right, improve=0.3440129, (0
missing)
##       parking_charges   < 73.5725   to the left,  improve=0.3059630, (0
missing)
##   Surrogate splits:
##       coop_condo        splits as  LR, agree=0.948, adj=0.842, (0 split)
##       monthly_cost      < 812       to the right, agree=0.948, adj=0.842,
(0 split)
##       approx_year_built < 1963.5    to the left,  agree=0.914, adj=0.737,
(0 split)
##       sq_footage        < 1143.5    to the right, agree=0.862, adj=0.579,
(0 split)
##       parking_charges   < 73.5725   to the left,  agree=0.793, adj=0.368,
(0 split)
##
## Node number 7: 26 observations,    complexity param=0.01487063
##   mean=664615.4, MSE=2.207662e+10
##   left son=14 (16 obs) right son=15 (10 obs)
##   Primary splits:
##       price_persqft          < 0.5451458 to the left,  improve=0.3412168,
(0 missing)
##       total_taxes            < 5018.9    to the left,  improve=0.3039461,
(0 missing)
##       parking_charges        < 141.8825  to the left,  improve=0.2528833,
(0 missing)
##       community_district_num < 27.5      to the left,  improve=0.2358061,
(0 missing)
##       num_half_bathrooms     < 0.945     to the right, improve=0.1572672,
(0 missing)
##   Surrogate splits:
```

```
##       total_taxes            < 5018.9    to the left,  agree=0.808,
adj=0.5, (0 split)
##       num_floors_in_building < 21        to the left,  agree=0.769,
adj=0.4, (0 split)
##       approx_year_built      < 1970.5    to the left,  agree=0.731,
adj=0.3, (0 split)
##       pct_tax_deductibl      < 35.69     to the right, agree=0.731,
adj=0.3, (0 split)
##       monthly_cost           < 2211      to the left,  agree=0.731,
adj=0.3, (0 split)
##
## Node number 8: 166 observations,    complexity param=0.007902237
##   mean=174793.3, MSE=3.347287e+09
##   left son=16 (122 obs) right son=17 (44 obs)
##   Primary splits:
##       price_persqft          < 0.4929002 to the left,  improve=0.1873079,
(0 missing)
##       walk_score             < 95.5      to the left,  improve=0.1524936,
(0 missing)
##       approx_year_built      < 1935.5    to the right, improve=0.1363928,
(0 missing)
##       parking_charges        < 118.57    to the left,  improve=0.1207486,
(0 missing)
##       community_district_num < 29.5      to the left,  improve=0.1175579,
(0 missing)
##   Surrogate splits:
##       parking_charges        < 130.89    to the left,  agree=0.843,
adj=0.409, (0 split)
##       walk_score             < 90.5      to the left,  agree=0.801,
adj=0.250, (0 split)
##       approx_year_built      < 1942.5    to the right, agree=0.789,
adj=0.205, (0 split)
##       num_floors_in_building < 8.64      to the left,  agree=0.789,
adj=0.205, (0 split)
##       total_taxes            < 2366.313  to the right, agree=0.789,
adj=0.205, (0 split)
##
## Node number 9: 118 observations,    complexity param=0.01515856
##   mean=269165.3, MSE=8.219769e+09
##   left son=18 (106 obs) right son=19 (12 obs)
##   Primary splits:
##       num_floors_in_building < 8.705     to the left,  improve=0.2058369,
(0 missing)
##       parking_charges        < 88.07     to the left,  improve=0.1792023,
(0 missing)
##       price_persqft          < 0.4516464 to the left,  improve=0.1596936,
(0 missing)
##       walk_score             < 91.5      to the left,  improve=0.1334127,
(0 missing)
##       monthly_cost           < 1048      to the left,  improve=0.1099628,
```

```
(0 missing)
##   Surrogate splits:
##       parking_charges   < 395.855   to the left,  agree=0.915, adj=0.167,
(0 split)
##       approx_year_built < 1964.5    to the left,  agree=0.907, adj=0.083,
(0 split)
##
## Node number 10: 32 observations,    complexity param=0.01878043
##   mean=389223.4, MSE=1.593423e+10
##   left son=20 (8 obs) right son=21 (24 obs)
##   Primary splits:
##       sq_footage      < 669       to the left,  improve=0.4851004, (0
missing)
##       num_total_rooms < 3.5       to the left,  improve=0.2499396, (0
missing)
##       parking_charges < 143.74    to the right, improve=0.2106907, (0
missing)
##       monthly_cost    < 297       to the left,  improve=0.1928689, (0
missing)
##       total_taxes     < 2503.925  to the left,  improve=0.1837103, (0
missing)
##   Surrogate splits:
##       num_bedrooms    < 0.5       to the left,  agree=0.812, adj=0.250, (0
split)
##       monthly_cost    < 177       to the left,  agree=0.812, adj=0.250, (0
split)
##       num_total_rooms < 2.5       to the left,  agree=0.781, adj=0.125, (0
split)
##       price_persqft   < 0.7280547 to the right, agree=0.781, adj=0.125, (0
split)
##
## Node number 11: 22 observations,    complexity param=0.005801889
##   mean=528727.3, MSE=1.257356e+10
##   left son=22 (15 obs) right son=23 (7 obs)
##   Primary splits:
##       approx_year_built      < 1987.5    to the left,  improve=0.2762457,
(0 missing)
##       community_district_num < 27        to the left,  improve=0.2757739,
(0 missing)
##       total_taxes            < 4034.245  to the left,  improve=0.2076008,
(0 missing)
##       parking_charges        < 140.855   to the left,  improve=0.1795759,
(0 missing)
##       num_floors_in_building < 12.5      to the right, improve=0.1745696,
(0 missing)
##   Surrogate splits:
##       price_persqft          < 0.6578675 to the left,  agree=0.909,
adj=0.714, (0 split)
##       community_district_num < 29        to the left,  agree=0.864,
adj=0.571, (0 split)
```

```
##       sq_footage             < 688.75    to the right, agree=0.864,
adj=0.571, (0 split)
##       pct_tax_deductibl      < 32.47     to the right, agree=0.818,
adj=0.429, (0 split)
##       total_taxes            < 4855.51   to the left,  agree=0.818,
adj=0.429, (0 split)
##
## Node number 12: 19 observations
##   mean=367844.7, MSE=8.995931e+09
##
## Node number 13: 39 observations,    complexity param=0.00618015
##   mean=536500, MSE=7.59891e+09
##   left son=26 (17 obs) right son=27 (22 obs)
##   Primary splits:
##       num_floors_in_building < 5.3075    to the left,  improve=0.27465670,
(0 missing)
##       pct_tax_deductibl      < 48.705    to the right, improve=0.18787370,
(0 missing)
##       total_taxes            < 2818.99   to the left,  improve=0.15733360,
(0 missing)
##       num_total_rooms        < 4.5       to the left,  improve=0.10917930,
(0 missing)
##       is_missing_total_taxes < 0.5       to the left,  improve=0.09994854,
(0 missing)
##   Surrogate splits:
##       total_taxes                < 3738.61   to the left,  agree=0.718,
adj=0.353, (0 split)
##       num_half_bathrooms         < 0.985     to the right, agree=0.692,
adj=0.294, (0 split)
##       pct_tax_deductibl          < 41.04     to the right, agree=0.692,
adj=0.294, (0 split)
##       walk_score                 < 85.5      to the left,  agree=0.692,
adj=0.294, (0 split)
##       is_missing_dining_room_type < 0.5       to the right, agree=0.692,
adj=0.294, (0 split)
##
## Node number 14: 16 observations
##   mean=596000, MSE=1.695512e+10
##
## Node number 15: 10 observations
##   mean=774400, MSE=1.068544e+10
##
## Node number 16: 122 observations,    complexity param=0.002727944
##   mean=159756, MSE=1.851336e+09
##   left son=32 (107 obs) right son=33 (15 obs)
##   Primary splits:
##       total_taxes       < 3829.475  to the left,  improve=0.15907340, (0
missing)
##       approx_year_built < 1957.5    to the right, improve=0.14962720, (0
missing)
```

```
##       walk_score        < 85.5      to the left,  improve=0.11881390, (0
missing)
##       sq_footage        < 671.99    to the left,  improve=0.07434716, (0
missing)
##       num_total_rooms   < 2.5       to the left,  improve=0.07263335, (0
missing)
##   Surrogate splits:
##       pct_tax_deductibl      < 39.585    to the right, agree=0.943,
adj=0.533, (0 split)
##       num_half_bathrooms     < 0.85      to the right, agree=0.934,
adj=0.467, (0 split)
##       num_floors_in_building < 1.5       to the right, agree=0.893,
adj=0.133, (0 split)
##
## Node number 17: 44 observations,    complexity param=0.005491657
##   mean=216487.7, MSE=5.129752e+09
##   left son=34 (36 obs) right son=35 (8 obs)
##   Primary splits:
##       sq_footage        < 776.6596  to the left,  improve=0.3204509, (0
missing)
##       pct_tax_deductibl < 48.555    to the right, improve=0.2670338, (0
missing)
##       total_taxes       < 2820.733  to the left,  improve=0.2273379, (0
missing)
##       monthly_cost      < 655.5     to the left,  improve=0.2191401, (0
missing)
##       num_total_rooms   < 2.5       to the left,  improve=0.2056785, (0
missing)
##   Surrogate splits:
##       monthly_cost           < 859       to the left,  agree=0.886,
adj=0.375, (0 split)
##       num_half_bathrooms     < 0.795     to the right, agree=0.864,
adj=0.250, (0 split)
##       pct_tax_deductibl      < 38.04     to the right, agree=0.864,
adj=0.250, (0 split)
##       total_taxes            < 3879.194  to the left,  agree=0.864,
adj=0.250, (0 split)
##       num_floors_in_building < 30        to the left,  agree=0.841,
adj=0.125, (0 split)
##
## Node number 18: 106 observations,    complexity param=0.006676428
##   mean=255325.5, MSE=6.709224e+09
##   left son=36 (61 obs) right son=37 (45 obs)
##   Primary splits:
##       parking_charges < 89.995    to the left,  improve=0.12364400, (0
missing)
##       price_persqft   < 0.4516464 to the left,  improve=0.08612148, (0
missing)
##       walk_score      < 91.5      to the left,  improve=0.08435061, (0
missing)
```

```
##       sq_footage      < 941.935   to the left,  improve=0.05881545, (0
missing)
##       monthly_cost    < 732.5     to the left,  improve=0.05726028, (0
missing)
##   Surrogate splits:
##       walk_score             < 86.5      to the left,  agree=0.868,
adj=0.689, (0 split)
##       price_persqft          < 0.4516464 to the left,  agree=0.868,
adj=0.689, (0 split)
##       community_district_num < 27.5      to the left,  agree=0.764,
adj=0.444, (0 split)
##       num_floors_in_building < 4.665     to the left,  agree=0.736,
adj=0.378, (0 split)
##       approx_year_built      < 1947.5    to the right, agree=0.698,
adj=0.289, (0 split)
##
## Node number 19: 12 observations
##   mean=391416.7, MSE=4.925576e+09
##
## Node number 20: 8 observations
##   mean=236943.8, MSE=5.856097e+09
##
## Node number 21: 24 observations,    complexity param=0.003773733
##   mean=439983.3, MSE=8.987341e+09
##   left son=42 (9 obs) right son=43 (15 obs)
##   Primary splits:
##       monthly_cost       < 313.5     to the left,  improve=0.23042830, (0
missing)
##       parking_charges    < 147.9042  to the right, improve=0.12270780, (0
missing)
##       pct_tax_deductibl  < 49.425    to the right, improve=0.10643120, (0
missing)
##       num_half_bathrooms < 1.055     to the right, improve=0.08634468, (0
missing)
##       total_taxes        < 2659.605  to the left,  improve=0.07811121, (0
missing)
##   Surrogate splits:
##       approx_year_built  < 1989.5    to the right, agree=0.833, adj=0.556,
(0 split)
##       num_half_bathrooms < 0.625     to the left,  agree=0.750, adj=0.333,
(0 split)
##       price_persqft      < 0.6635951 to the right, agree=0.750, adj=0.333,
(0 split)
##       parking_charges    < 141.0575  to the left,  agree=0.708, adj=0.222,
(0 split)
##       pct_tax_deductibl  < 37.66     to the left,  agree=0.708, adj=0.222,
(0 split)
##
## Node number 22: 15 observations
##   mean=488466.7, MSE=6.202782e+09
```

```
##
## Node number 23: 7 observations
##   mean=615000, MSE=1.530886e+10
##
## Node number 26: 17 observations
##   mean=484529.4, MSE=7.546484e+09
##
## Node number 27: 22 observations,    complexity param=0.0009341672
##   mean=576659.1, MSE=3.939577e+09
##   left son=54 (8 obs) right son=55 (14 obs)
##   Primary splits:
##       sq_footage         < 969.295   to the left,  improve=0.14195790, (0
missing)
##       num_half_bathrooms < 0.44      to the left,  improve=0.06590791, (0
missing)
##       pct_tax_deductibl  < 34.33     to the right, improve=0.05286468, (0
missing)
##       total_taxes        < 4044.115  to the right, improve=0.04918481, (0
missing)
##       monthly_cost       < 521.5     to the left,  improve=0.04236894, (0
missing)
##   Surrogate splits:
##       num_total_rooms        < 4.5       to the left,  agree=0.818,
adj=0.500, (0 split)
##       pct_tax_deductibl      < 33.81     to the left,  agree=0.773,
adj=0.375, (0 split)
##       walk_score             < 95.5      to the right, agree=0.773,
adj=0.375, (0 split)
##       is_missing_total_taxes < 0.5       to the left,  agree=0.773,
adj=0.375, (0 split)
##       approx_year_built      < 2008.5    to the right, agree=0.727,
adj=0.250, (0 split)
##
## Node number 32: 107 observations,    complexity param=0.002182803
##   mean=153330.6, MSE=1.462748e+09
##   left son=64 (29 obs) right son=65 (78 obs)
##   Primary splits:
##       approx_year_built            < 1957.5    to the right,
improve=0.18368280, (0 missing)
##       walk_score                   < 84        to the left,
improve=0.11396640, (0 missing)
##       total_taxes                  < 2416.155  to the left,
improve=0.08060385, (0 missing)
##       num_total_rooms              < 2.5       to the left,
improve=0.07544029, (0 missing)
##       is_missing_pct_tax_deductibl < 0.5       to the left,
improve=0.07128177, (0 missing)
##   Surrogate splits:
##       num_floors_in_building < 6.24      to the right, agree=0.757,
adj=0.103, (0 split)
```

```
##       parking_charges        < 153.5617  to the right, agree=0.748,
adj=0.069, (0 split)
##       monthly_cost           < 404.5     to the left,  agree=0.748,
adj=0.069, (0 split)
##       price_persqft          < 0.3485595 to the left,  agree=0.748,
adj=0.069, (0 split)
##       community_district_num < 24.5      to the left,  agree=0.738,
adj=0.034, (0 split)
##
## Node number 33: 15 observations
##   mean=205589.9, MSE=2.228011e+09
##
## Node number 34: 36 observations,    complexity param=0.002336409
##   mean=197375, MSE=3.236741e+09
##   left son=68 (20 obs) right son=69 (16 obs)
##   Primary splits:
##       total_taxes      < 2528.09   to the left,  improve=0.2640860, (0
missing)
##       dining_room_type splits as  L-R-R, improve=0.2162890, (0 missing)
##       kitchen_type     splits as  LRRL, improve=0.1890938, (0 missing)
##       num_total_rooms  < 2.5       to the left,  improve=0.1875898, (0
missing)
##       sq_footage       < 624.2767  to the left,  improve=0.1740703, (0
missing)
##   Surrogate splits:
##       monthly_cost      < 581       to the left,  agree=0.861, adj=0.688,
(0 split)
##       sq_footage        < 723.91    to the left,  agree=0.833, adj=0.625,
(0 split)
##       parking_charges   < 139.635   to the right, agree=0.806, adj=0.562,
(0 split)
##       pct_tax_deductibl < 49.18     to the right, agree=0.806, adj=0.562,
(0 split)
##       dining_room_type  splits as  L-R-R, agree=0.722, adj=0.375, (0
split)
##
## Node number 35: 8 observations
##   mean=302495, MSE=4.607215e+09
##
## Node number 36: 61 observations,    complexity param=0.00494991
##   mean=230587.5, MSE=4.122257e+09
##   left son=72 (13 obs) right son=73 (48 obs)
##   Primary splits:
##       community_district_num < 26.5      to the right, improve=0.2592624,
(0 missing)
##       pct_tax_deductibl      < 48.495    to the right, improve=0.1883041,
(0 missing)
##       total_taxes            < 2988.033  to the left,  improve=0.1553637,
(0 missing)
##       num_total_rooms        < 4.5       to the left,  improve=0.1337204,
```

```
(0 missing)
##       monthly_cost           < 732.5     to the left,  improve=0.1291645,
(0 missing)
##   Surrogate splits:
##       approx_year_built < 1962      to the right, agree=0.836, adj=0.231,
(0 split)
##       total_taxes       < 2778.18   to the left,  agree=0.820, adj=0.154,
(0 split)
##       monthly_cost      < 1305.5    to the right, agree=0.820, adj=0.154,
(0 split)
##       price_persqft     < 0.4487844 to the right, agree=0.820, adj=0.154,
(0 split)
##       walk_score        < 87.5      to the right, agree=0.803, adj=0.077,
(0 split)
##
## Node number 37: 45 observations,    complexity param=0.005607514
##   mean=288859.2, MSE=8.26194e+09
##   left son=74 (35 obs) right son=75 (10 obs)
##   Primary splits:
##       monthly_cost          < 905.5     to the left,  improve=0.1986474,
(0 missing)
##       is_missing_sq_footage < 0.5       to the left,  improve=0.1505807,
(0 missing)
##       kitchen_type          splits as  LLLR, improve=0.1309333, (0
missing)
##       num_bedrooms          < 1.5       to the left,  improve=0.1217885,
(0 missing)
##       num_half_bathrooms    < 0.985     to the right, improve=0.0830978,
(0 missing)
##   Surrogate splits:
##       price_persqft < 0.4106358 to the right, agree=0.844, adj=0.3, (0
split)
##       sq_footage    < 1225      to the left,  agree=0.822, adj=0.2, (0
split)
##
## Node number 42: 9 observations
##   mean=381233.3, MSE=7.553138e+09
##
## Node number 43: 15 observations
##   mean=475233.3, MSE=6.534362e+09
##
## Node number 54: 8 observations
##   mean=545375, MSE=2.134734e+09
##
## Node number 55: 14 observations
##   mean=594535.7, MSE=4.092088e+09
##
## Node number 64: 29 observations,    complexity param=0.0005192759
##   mean=126448.3, MSE=1.032816e+09
##   left son=128 (9 obs) right son=129 (20 obs)
```

```
##   Primary splits:
##       sq_footage             < 732.7425  to the left,  improve=0.22834120,
(0 missing)
##       community_district_num < 26.5      to the right, improve=0.20134940,
(0 missing)
##       monthly_cost           < 560.5     to the left,  improve=0.14040960,
(0 missing)
##       approx_year_built      < 1964      to the right, improve=0.12082800,
(0 missing)
##       total_taxes            < 2635.46   to the left,  improve=0.07210595,
(0 missing)
##   Surrogate splits:
##       total_taxes            < 2556.055  to the left,  agree=0.897,
adj=0.667, (0 split)
##       num_total_rooms        < 2.5       to the left,  agree=0.862,
adj=0.556, (0 split)
##       monthly_cost           < 573       to the left,  agree=0.862,
adj=0.556, (0 split)
##       num_bedrooms           < 0.5       to the left,  agree=0.828,
adj=0.444, (0 split)
##       num_floors_in_building < 4.7275    to the left,  agree=0.759,
adj=0.222, (0 split)
##
## Node number 65: 78 observations,    complexity param=0.001424518
##   mean=163325.3, MSE=1.254018e+09
##   left son=130 (51 obs) right son=131 (27 obs)
##   Primary splits:
##       walk_score                   < 85.5      to the left,
improve=0.19181230, (0 missing)
##       community_district_num       < 28.5      to the right,
improve=0.10639520, (0 missing)
##       sq_footage                   < 702.19    to the left,
improve=0.09296473, (0 missing)
##       num_floors_in_building       < 6.5       to the left,
improve=0.07363263, (0 missing)
##       is_missing_pct_tax_deductibl < 0.5       to the left,
improve=0.07358039, (0 missing)
##   Surrogate splits:
##       parking_charges        < 99.45     to the left,  agree=0.885,
adj=0.667, (0 split)
##       num_floors_in_building < 5.71      to the left,  agree=0.821,
adj=0.481, (0 split)
##       price_persqft          < 0.4698517 to the left,  agree=0.756,
adj=0.296, (0 split)
##       sq_footage             < 560.5     to the right, agree=0.718,
adj=0.185, (0 split)
##       num_bedrooms           < 0.5       to the right, agree=0.705,
adj=0.148, (0 split)
##
## Node number 68: 20 observations,    complexity param=0.0007780139
```

```
##   mean=171225, MSE=2.145562e+09
##   left son=136 (13 obs) right son=137 (7 obs)
##   Primary splits:
##       kitchen_type      splits as  -LRL, improve=0.2387940, (0 missing)
##       pct_tax_deductibl < 49.5      to the left,  improve=0.2281871, (0
missing)
##       total_taxes       < 2289.13   to the right, improve=0.2036933, (0
missing)
##       num_total_rooms   < 2.5       to the left,  improve=0.1678118, (0
missing)
##       sq_footage        < 578.1933  to the left,  improve=0.1138223, (0
missing)
##   Surrogate splits:
##       approx_year_built      < 1938      to the right, agree=0.75,
adj=0.286, (0 split)
##       community_district_num < 29        to the left,  agree=0.75,
adj=0.286, (0 split)
##       num_half_bathrooms     < 1.095     to the left,  agree=0.75,
adj=0.286, (0 split)
##       walk_score             < 93.5      to the left,  agree=0.75,
adj=0.286, (0 split)
##       dining_room_type       splits as  L---R, agree=0.70, adj=0.143, (0
split)
##
## Node number 69: 16 observations
##   mean=230062.5, MSE=2.677465e+09
##
## Node number 72: 13 observations
##   mean=167769.2, MSE=2.398947e+09
##
## Node number 73: 48 observations,    complexity param=0.001881432
##   mean=247600.8, MSE=3.230788e+09
##   left son=146 (20 obs) right son=147 (28 obs)
##   Primary splits:
##       num_total_rooms    < 4.5       to the left,  improve=0.15978860, (0
missing)
##       total_taxes        < 4153.965  to the left,  improve=0.13981400, (0
missing)
##       monthly_cost       < 721       to the left,  improve=0.12308680, (0
missing)
##       num_half_bathrooms < 0.915     to the left,  improve=0.10145040, (0
missing)
##       num_bedrooms       < 2.5       to the left,  improve=0.09582418, (0
missing)
##   Surrogate splits:
##       sq_footage         < 900.245   to the left,  agree=0.833, adj=0.60,
(0 split)
##       num_half_bathrooms < 0.975     to the left,  agree=0.812, adj=0.55,
(0 split)
##       price_persqft      < 0.3910969 to the left,  agree=0.771, adj=0.45,
```

```
(0 split)
##       pct_tax_deductibl  < 40.255    to the left,  agree=0.708, adj=0.30,
(0 split)
##       monthly_cost       < 699       to the left,  agree=0.708, adj=0.30,
(0 split)
##
## Node number 74: 35 observations,    complexity param=0.004412636
##   mean=267204.7, MSE=7.662762e+09
##   left son=148 (9 obs) right son=149 (26 obs)
##   Primary splits:
##       num_bedrooms           < 1.5       to the left,  improve=0.21669650,
(0 missing)
##       is_missing_sq_footage  < 0.5       to the left,  improve=0.16516260,
(0 missing)
##       pets_allowed           < 0.5       to the left,  improve=0.07508865,
(0 missing)
##       pct_tax_deductibl      < 41.965    to the left,  improve=0.07385266,
(0 missing)
##       community_district_num < 27        to the right, improve=0.07369889,
(0 missing)
##   Surrogate splits:
##       num_total_rooms    < 3.5       to the left,  agree=0.886, adj=0.556,
(0 split)
##       sq_footage         < 856.07    to the left,  agree=0.857, adj=0.444,
(0 split)
##       total_taxes        < 2429.22   to the left,  agree=0.829, adj=0.333,
(0 split)
##       approx_year_built  < 1937.5    to the left,  agree=0.800, adj=0.222,
(0 split)
##       num_half_bathrooms < 1.25      to the right, agree=0.800, adj=0.222,
(0 split)
##
## Node number 75: 10 observations
##   mean=364650, MSE=2.973602e+09
##
## Node number 128: 9 observations
##   mean=103555.6, MSE=1.108691e+09
##
## Node number 129: 20 observations,    complexity param=0.0004022024
##   mean=136750, MSE=6.567125e+08
##   left son=258 (7 obs) right son=259 (13 obs)
##   Primary splits:
##       pct_tax_deductibl < 46.56     to the left,  improve=0.4033170, (0
missing)
##       total_taxes       < 3000.305  to the right, improve=0.4033170, (0
missing)
##       sq_footage        < 753.9375  to the right, improve=0.2460066, (0
missing)
##       approx_year_built < 1964      to the right, improve=0.2241947, (0
missing)
```

```
##       walk_score        < 76.5      to the right, improve=0.2108415, (0
missing)
##   Surrogate splits:
##       total_taxes       < 3000.305  to the right, agree=1.00, adj=1.000,
(0 split)
##       price_persqft     < 0.3831208 to the left,  agree=0.85, adj=0.571,
(0 split)
##       approx_year_built < 1964      to the right, agree=0.80, adj=0.429,
(0 split)
##       monthly_cost      < 742.5     to the right, agree=0.80, adj=0.429,
(0 split)
##       dining_room_type  splits as  R-R-L, agree=0.75, adj=0.286, (0 split)
##
## Node number 130: 51 observations,    complexity param=0.000776772
##   mean=152040.7, MSE=7.061644e+08
##   left son=260 (9 obs) right son=261 (42 obs)
##   Primary splits:
##       community_district_num < 28.5      to the right, improve=0.2840696,
(0 missing)
##       parking_charges        < 95.5075   to the right, improve=0.2302079,
(0 missing)
##       num_half_bathrooms     < 0.955     to the right, improve=0.2045244,
(0 missing)
##       pets_allowed           < 0.5       to the left,  improve=0.1461233,
(0 missing)
##       pct_tax_deductibl      < 49.2      to the right, improve=0.1295276,
(0 missing)
##   Surrogate splits:
##       num_half_bathrooms < 1.005     to the right, agree=0.882, adj=0.333,
(0 split)
##       pct_tax_deductibl  < 50.84     to the right, agree=0.882, adj=0.333,
(0 split)
##       total_taxes        < 2111.518  to the left,  agree=0.882, adj=0.333,
(0 split)
##       price_persqft      < 0.4695231 to the right, agree=0.882, adj=0.333,
(0 split)
##       parking_charges    < 126.165   to the right, agree=0.863, adj=0.222,
(0 split)
##
## Node number 131: 27 observations,    complexity param=0.001399473
##   mean=184640.7, MSE=1.593971e+09
##   left son=262 (8 obs) right son=263 (19 obs)
##   Primary splits:
##       sq_footage                         < 691.845   to the left,
improve=0.42827970, (0 missing)
##       monthly_cost                       < 551.5     to the left,
improve=0.18372920, (0 missing)
##       walk_score                         < 92.5      to the right,
improve=0.17181720, (0 missing)
##       total_taxes                        < 2604.992  to the left,
```

```
improve=0.09958519, (0 missing)
##       is_missing_num_floors_in_building < 0.5       to the left,
improve=0.08229344, (0 missing)
##   Surrogate splits:
##       monthly_cost           < 541.5     to the left,  agree=0.926,
adj=0.75, (0 split)
##       num_bedrooms           < 0.5       to the left,  agree=0.852,
adj=0.50, (0 split)
##       num_total_rooms        < 2.5       to the left,  agree=0.852,
adj=0.50, (0 split)
##       community_district_num < 18        to the left,  agree=0.778,
adj=0.25, (0 split)
##       parking_charges        < 149.785   to the right, agree=0.778,
adj=0.25, (0 split)
##
## Node number 136: 13 observations
##   mean=154615.4, MSE=1.460237e+09
##
## Node number 137: 7 observations
##   mean=202071.4, MSE=1.954459e+09
##
## Node number 146: 20 observations,    complexity param=0.001057001
##   mean=220717, MSE=2.573899e+09
##   left son=292 (7 obs) right son=293 (13 obs)
##   Primary splits:
##       monthly_cost               < 671       to the left,
improve=0.2704339, (0 missing)
##       is_missing_parking_charges < 0.5       to the left,
improve=0.1964711, (0 missing)
##       walk_score                 < 76.5      to the left,
improve=0.1666119, (0 missing)
##       parking_charges            < 61.47     to the right,
improve=0.1408563, (0 missing)
##       pct_tax_deductibl          < 41.07     to the right,
improve=0.1346381, (0 missing)
##   Surrogate splits:
##       pct_tax_deductibl      < 45.83     to the right, agree=0.95,
adj=0.857, (0 split)
##       total_taxes            < 3361.625  to the left,  agree=0.95,
adj=0.857, (0 split)
##       price_persqft          < 0.4134622 to the right, agree=0.80,
adj=0.429, (0 split)
##       kitchen_type           splits as  -LRR, agree=0.75, adj=0.286, (0
split)
##       community_district_num < 25.5      to the right, agree=0.70,
adj=0.143, (0 split)
##
## Node number 147: 28 observations,    complexity param=0.001033753
##   mean=266803.6, MSE=2.815006e+09
##   left son=294 (8 obs) right son=295 (20 obs)
```

```
##   Primary splits:
##       num_floors_in_building < 3.475     to the right, improve=0.17273750,
(0 missing)
##       pct_tax_deductibl      < 48.05     to the left,  improve=0.15883810,
(0 missing)
##       sq_footage             < 895.625   to the right, improve=0.12603770,
(0 missing)
##       total_taxes            < 3199.02   to the right, improve=0.08125236,
(0 missing)
##       walk_score             < 74        to the right, improve=0.07930895,
(0 missing)
##   Surrogate splits:
##       approx_year_built      < 1951.5    to the right, agree=0.857,
adj=0.500, (0 split)
##       community_district_num < 24.5      to the left,  agree=0.786,
adj=0.250, (0 split)
##       walk_score             < 79.5      to the right, agree=0.786,
adj=0.250, (0 split)
##       parking_charges        < 76.8975   to the right, agree=0.750,
adj=0.125, (0 split)
##       sq_footage             < 977.07    to the right, agree=0.750,
adj=0.125, (0 split)
##
## Node number 148: 9 observations
##   mean=197944.4, MSE=2.101136e+09
##
## Node number 149: 26 observations,    complexity param=0.003050957
##   mean=291179.4, MSE=7.352661e+09
##   left son=298 (17 obs) right son=299 (9 obs)
##   Primary splits:
##       monthly_cost           < 772       to the right, improve=0.2101964,
(0 missing)
##       community_district_num < 27        to the right, improve=0.1336729,
(0 missing)
##       pets_allowed           < 0.5       to the left,  improve=0.1263586,
(0 missing)
##       pct_tax_deductibl      < 41.915    to the left,  improve=0.1158970,
(0 missing)
##       is_missing_sq_footage  < 0.5       to the left,  improve=0.1007128,
(0 missing)
##   Surrogate splits:
##       community_district_num < 21.5      to the right, agree=0.769,
adj=0.333, (0 split)
##       parking_charges        < 94.7775   to the right, agree=0.769,
adj=0.333, (0 split)
##       sq_footage             < 874.47    to the right, agree=0.769,
adj=0.333, (0 split)
##       walk_score             < 96.5      to the left,  agree=0.769,
adj=0.333, (0 split)
##       coop_condo             splits as  LR, agree=0.731, adj=0.222, (0
```

```
split)
##
## Node number 258: 7 observations
##   mean=114571.4, MSE=7.662449e+08
##
## Node number 259: 13 observations
##   mean=148692.3, MSE=1.902515e+08
##
## Node number 260: 9 observations
##   mean=121444.4, MSE=2.766358e+08
##
## Node number 261: 42 observations,    complexity param=0.0002183995
##   mean=158597.1, MSE=5.546208e+08
##   left son=522 (23 obs) right son=523 (19 obs)
##   Primary splits:
##       num_half_bathrooms     < 0.945     to the right, improve=0.11666630,
(0 missing)
##       walk_score             < 77.5      to the right, improve=0.09660561,
(0 missing)
##       is_missing_sq_footage  < 0.5       to the left,  improve=0.08132811,
(0 missing)
##       monthly_cost           < 599.5     to the right, improve=0.06106717,
(0 missing)
##       is_missing_total_taxes < 0.5       to the right, improve=0.06073954,
(0 missing)
##   Surrogate splits:
##       parking_charges   < 67.6675   to the right, agree=0.857, adj=0.684,
(0 split)
##       price_persqft     < 0.3922014 to the right, agree=0.833, adj=0.632,
(0 split)
##       sq_footage        < 732.9825  to the left,  agree=0.762, adj=0.474,
(0 split)
##       total_taxes       < 2775.515  to the left,  agree=0.714, adj=0.368,
(0 split)
##       pct_tax_deductibl < 39.85     to the right, agree=0.667, adj=0.263,
(0 split)
##
## Node number 262: 8 observations
##   mean=144375, MSE=3.219844e+08
##
## Node number 263: 19 observations
##   mean=201594.7, MSE=1.15944e+09
##
## Node number 292: 7 observations
##   mean=184762.9, MSE=8.277374e+08
##
## Node number 293: 13 observations
##   mean=240076.9, MSE=2.443263e+09
##
## Node number 294: 8 observations
```

```
##   mean=231937.5, MSE=1.010402e+09
##
## Node number 295: 20 observations,    complexity param=0.0006390631
##   mean=280750, MSE=2.856088e+09
##   left son=590 (12 obs) right son=591 (8 obs)
##   Primary splits:
##       num_half_bathrooms < 0.995     to the left,  improve=0.14734970, (0
missing)
##       monthly_cost       < 838       to the right, improve=0.12974480, (0
missing)
##       pct_tax_deductibl  < 47.895    to the left,  improve=0.08173428, (0
missing)
##       parking_charges    < 56.8875   to the left,  improve=0.07300798, (0
missing)
##       sq_footage         < 895.625   to the right, improve=0.06577646, (0
missing)
##   Surrogate splits:
##       sq_footage        < 929.845   to the left,  agree=0.85, adj=0.625,
(0 split)
##       approx_year_built < 1952.5    to the left,  agree=0.80, adj=0.500,
(0 split)
##       parking_charges   < 45.53     to the right, agree=0.80, adj=0.500,
(0 split)
##       price_persqft     < 0.4281089 to the left,  agree=0.80, adj=0.500,
(0 split)
##       monthly_cost      < 952.5     to the left,  agree=0.75, adj=0.375,
(0 split)
##
## Node number 298: 17 observations
##   mean=262575.1, MSE=6.041012e+09
##
## Node number 299: 9 observations
##   mean=345209.8, MSE=5.365433e+09
##
## Node number 522: 23 observations,    complexity param=0.0002183995
##   mean=151286, MSE=7.013909e+08
##   left son=1044 (16 obs) right son=1045 (7 obs)
##   Primary splits:
##       monthly_cost           < 599.5     to the right, improve=0.1881536,
(0 missing)
##       is_missing_sq_footage  < 0.5       to the left,  improve=0.1651616,
(0 missing)
##       sq_footage             < 718.605   to the right, improve=0.1619020,
(0 missing)
##       num_floors_in_building < 3.655     to the left,  improve=0.1251356,
(0 missing)
##       num_half_bathrooms     < 0.975     to the left,  improve=0.1031153,
(0 missing)
##   Surrogate splits:
##       total_taxes            < 2555.94   to the right, agree=0.870,
```

```
adj=0.571, (0 split)
##       dining_room_type       splits as  L-R-R, agree=0.826, adj=0.429, (0
split)
##       price_persqft          < 0.4573344 to the left,  agree=0.826,
adj=0.429, (0 split)
##       community_district_num < 26.5      to the left,  agree=0.739,
adj=0.143, (0 split)
##       num_floors_in_building < 4.005     to the left,  agree=0.739,
adj=0.143, (0 split)
##
## Node number 523: 19 observations
##   mean=167447.4, MSE=2.339183e+08
##
## Node number 590: 12 observations
##   mean=264000, MSE=1.357333e+09
##
## Node number 591: 8 observations
##   mean=305875, MSE=4.052109e+09
##
## Node number 1044: 16 observations
##   mean=143687.5, MSE=5.285898e+08
##
## Node number 1045: 7 observations
##   mean=168653.9, MSE=6.627517e+08
```

```
yhat2 <- predict(mod2, Xtest)
e2 <- yhat2 - Ytest
sqrt(sum(e2^2)/106)
```

```
## [1] 103078.1
```

Tuning

```
mod3 <- rpart(
  formula = Ytrain ~ .,
  data    = Xtrain,
  method  = "anova",
  control = list(minsplit = 10, maxdepth = 12, xval = 10)
)
yhat3 <- predict(mod3, Xtest)
summary(mod3)
```

```
## Call:
## rpart(formula = Ytrain ~ ., data = Xtrain, method = "anova",
##     control = list(minsplit = 10, maxdepth = 12, xval = 10))
##   n= 422
##
##            CP nsplit rel error    xerror       xstd
## 1  0.42055632      0 1.0000000 1.0060020 0.07564038
## 2  0.18550984      1 0.5794437 0.6916319 0.06685520
## 3  0.04663919      2 0.3939338 0.4535066 0.04200201
```

```
## 4  0.04582955      3 0.3472947 0.4130265 0.03798872
## 5  0.02759198      4 0.3014651 0.3662018 0.03528808
## 6  0.01926389      5 0.2738731 0.3389620 0.03424167
## 7  0.01878043      6 0.2546092 0.3295965 0.03443944
## 8  0.01515856      7 0.2358288 0.3207510 0.03215055
## 9  0.01487063      8 0.2206703 0.3148164 0.03203113
## 10 0.01000000      9 0.2057996 0.2889764 0.02949261
##
## Variable importance
##     num_full_bathrooms             sq_footage           monthly_cost
##                     19                     17                     11
##          price_persqft            total_taxes        num_total_rooms
##                     10                      9                      8
##             coop_condo      approx_year_built     num_half_bathrooms
##                      7                      7                      4
##           num_bedrooms num_floors_in_building      pct_tax_deductibl
##                      3                      3                      2
##        parking_charges
##                      1
##
## Node number 1: 422 observations,    complexity param=0.4205563
##   mean=308191.7, MSE=3.121006e+10
##   left son=2 (338 obs) right son=3 (84 obs)
##   Primary splits:
##       num_full_bathrooms < 1.5       to the left,  improve=0.4205563, (0
missing)
##       price_persqft      < 0.5527845 to the left,  improve=0.3922844, (0
missing)
##       coop_condo         splits as  LR, improve=0.3754617, (0 missing)
##       approx_year_built  < 1970.5    to the left,  improve=0.3463094, (0
missing)
##       sq_footage         < 969.225   to the left,  improve=0.3139306, (0
missing)
##   Surrogate splits:
##       sq_footage         < 1133.495  to the left,  agree=0.917, adj=0.583,
(0 split)
##       total_taxes        < 4391.615  to the left,  agree=0.870, adj=0.345,
(0 split)
##       num_total_rooms    < 5.5       to the left,  agree=0.860, adj=0.298,
(0 split)
##       monthly_cost       < 1274.5    to the left,  agree=0.844, adj=0.214,
(0 split)
##       num_half_bathrooms < 0.455     to the right, agree=0.839, adj=0.190,
(0 split)
##
## Node number 2: 338 observations,    complexity param=0.1855098
##   mean=251077.9, MSE=1.663718e+10
##   left son=4 (284 obs) right son=5 (54 obs)
##   Primary splits:
##       price_persqft     < 0.5541679 to the left,  improve=0.4344879, (0
```

```
missing)
##       coop_condo        splits as  LR, improve=0.3876538, (0 missing)
##       approx_year_built < 1970.5    to the left,  improve=0.3067765, (0
missing)
##       sq_footage        < 857.955   to the left,  improve=0.1567627, (0
missing)
##       total_taxes       < 4198.69   to the left,  improve=0.1488824, (0
missing)
##   Surrogate splits:
##       coop_condo        splits as  LR, agree=0.956, adj=0.722, (0 split)
##       approx_year_built < 1970.5    to the left,  agree=0.944, adj=0.648,
(0 split)
##       monthly_cost      < 390.5     to the right, agree=0.902, adj=0.389,
(0 split)
##       pct_tax_deductibl < 34.945    to the right, agree=0.864, adj=0.148,
(0 split)
##       total_taxes       < 4463.51   to the left,  agree=0.861, adj=0.130,
(0 split)
##
## Node number 3: 84 observations,    complexity param=0.04582955
##   mean=538006.5, MSE=2.390811e+10
##   left son=6 (58 obs) right son=7 (26 obs)
##   Primary splits:
##       sq_footage             < 1368.09   to the left,  improve=0.3005579,
(0 missing)
##       price_persqft          < 0.4700775 to the left,  improve=0.2564064,
(0 missing)
##       num_floors_in_building < 14.5      to the left,  improve=0.2206020,
(0 missing)
##       parking_charges        < 69.735    to the left,  improve=0.2094910,
(0 missing)
##       total_taxes            < 4976.01   to the left,  improve=0.2043087,
(0 missing)
##   Surrogate splits:
##       monthly_cost           < 1461.5    to the left,  agree=0.857,
adj=0.538, (0 split)
##       num_bedrooms           < 2.5       to the left,  agree=0.833,
adj=0.462, (0 split)
##       num_floors_in_building < 11.525    to the left,  agree=0.798,
adj=0.346, (0 split)
##       num_total_rooms        < 6.5       to the left,  agree=0.774,
adj=0.269, (0 split)
##       total_taxes            < 4787.5    to the left,  agree=0.762,
adj=0.231, (0 split)
##
## Node number 4: 284 observations,    complexity param=0.04663919
##   mean=214004.2, MSE=7.534685e+09
##   left son=8 (166 obs) right son=9 (118 obs)
##   Primary splits:
##       sq_footage      < 800.5     to the left,  improve=0.2870613, (0
```

```
missing)
##       monthly_cost    < 761.5     to the left,  improve=0.2363109, (0
missing)
##       num_bedrooms    < 1.5       to the left,  improve=0.2009774, (0
missing)
##       num_total_rooms < 4.5       to the left,  improve=0.1218869, (0
missing)
##       price_persqft   < 0.5121664 to the left,  improve=0.1065049, (0
missing)
##   Surrogate splits:
##       num_bedrooms     < 1.5       to the left,  agree=0.891, adj=0.737,
(0 split)
##       num_total_rooms  < 3.5       to the left,  agree=0.835, adj=0.602,
(0 split)
##       monthly_cost     < 761.5     to the left,  agree=0.803, adj=0.525,
(0 split)
##       total_taxes      < 2831.22   to the left,  agree=0.704, adj=0.288,
(0 split)
##       dining_room_type splits as  LLR-R, agree=0.648, adj=0.153, (0 split)
##
## Node number 5: 54 observations,    complexity param=0.01926389
##   mean=446058.3, MSE=1.926355e+10
##   left son=10 (32 obs) right son=11 (22 obs)
##   Primary splits:
##       num_floors_in_building < 7.125     to the left,  improve=0.2439052,
(0 missing)
##       monthly_cost           < 304       to the left,  improve=0.2437085,
(0 missing)
##       total_taxes            < 2503.925  to the left,  improve=0.2306287,
(0 missing)
##       num_half_bathrooms     < 1.015     to the right, improve=0.2085708,
(0 missing)
##       sq_footage             < 669       to the left,  improve=0.2063172,
(0 missing)
##   Surrogate splits:
##       monthly_cost       < 401.5     to the left,  agree=0.796, adj=0.500,
(0 split)
##       num_half_bathrooms < 1.005     to the right, agree=0.741, adj=0.364,
(0 split)
##       pct_tax_deductibl  < 47.865    to the right, agree=0.722, adj=0.318,
(0 split)
##       total_taxes        < 4255.025  to the left,  agree=0.722, adj=0.318,
(0 split)
##       parking_charges    < 135.86    to the right, agree=0.685, adj=0.227,
(0 split)
##
## Node number 6: 58 observations,    complexity param=0.02759198
##   mean=481250.9, MSE=1.432215e+10
##   left son=12 (19 obs) right son=13 (39 obs)
##   Primary splits:
```

```
##       price_persqft     < 0.5087043 to the left,  improve=0.4374757, (0
missing)
##       approx_year_built < 1966.5    to the left,  improve=0.4026342, (0
missing)
##       coop_condo        splits as  LR, improve=0.3440129, (0 missing)
##       monthly_cost      < 812       to the right, improve=0.3440129, (0
missing)
##       parking_charges   < 73.5725   to the left,  improve=0.3059630, (0
missing)
##   Surrogate splits:
##       coop_condo        splits as  LR, agree=0.948, adj=0.842, (0 split)
##       monthly_cost      < 812       to the right, agree=0.948, adj=0.842,
(0 split)
##       approx_year_built < 1963.5    to the left,  agree=0.914, adj=0.737,
(0 split)
##       sq_footage        < 1143.5    to the right, agree=0.862, adj=0.579,
(0 split)
##       parking_charges   < 73.5725   to the left,  agree=0.793, adj=0.368,
(0 split)
##
## Node number 7: 26 observations,    complexity param=0.01487063
##   mean=664615.4, MSE=2.207662e+10
##   left son=14 (16 obs) right son=15 (10 obs)
##   Primary splits:
##       price_persqft          < 0.5451458 to the left,  improve=0.3412168,
(0 missing)
##       total_taxes            < 5018.9    to the left,  improve=0.3039461,
(0 missing)
##       parking_charges        < 141.8825  to the left,  improve=0.2528833,
(0 missing)
##       community_district_num < 27.5      to the left,  improve=0.2358061,
(0 missing)
##       num_half_bathrooms     < 0.945     to the right, improve=0.1572672,
(0 missing)
##   Surrogate splits:
##       total_taxes            < 5018.9    to the left,  agree=0.808,
adj=0.5, (0 split)
##       num_floors_in_building < 21        to the left,  agree=0.769,
adj=0.4, (0 split)
##       approx_year_built      < 1970.5    to the left,  agree=0.731,
adj=0.3, (0 split)
##       pct_tax_deductibl      < 35.69     to the right, agree=0.731,
adj=0.3, (0 split)
##       monthly_cost           < 2211      to the left,  agree=0.731,
adj=0.3, (0 split)
##
## Node number 8: 166 observations
##   mean=174793.3, MSE=3.347287e+09
##
## Node number 9: 118 observations,    complexity param=0.01515856
```

```
##   mean=269165.3, MSE=8.219769e+09
##   left son=18 (106 obs) right son=19 (12 obs)
##   Primary splits:
##       num_floors_in_building < 8.705     to the left,  improve=0.2058369,
(0 missing)
##       parking_charges        < 88.07     to the left,  improve=0.1792023,
(0 missing)
##       price_persqft          < 0.4516464 to the left,  improve=0.1596936,
(0 missing)
##       walk_score             < 91.5      to the left,  improve=0.1334127,
(0 missing)
##       monthly_cost           < 1048      to the left,  improve=0.1099628,
(0 missing)
##   Surrogate splits:
##       parking_charges   < 395.855   to the left,  agree=0.915, adj=0.167,
(0 split)
##       approx_year_built < 1964.5    to the left,  agree=0.907, adj=0.083,
(0 split)
##
## Node number 10: 32 observations,    complexity param=0.01878043
##   mean=389223.4, MSE=1.593423e+10
##   left son=20 (8 obs) right son=21 (24 obs)
##   Primary splits:
##       sq_footage      < 669       to the left,  improve=0.4851004, (0
missing)
##       num_total_rooms < 3.5       to the left,  improve=0.2499396, (0
missing)
##       parking_charges < 143.74    to the right, improve=0.2106907, (0
missing)
##       monthly_cost    < 297       to the left,  improve=0.1928689, (0
missing)
##       total_taxes     < 2503.925  to the left,  improve=0.1837103, (0
missing)
##   Surrogate splits:
##       num_bedrooms    < 0.5       to the left,  agree=0.812, adj=0.250, (0
split)
##       monthly_cost    < 177       to the left,  agree=0.812, adj=0.250, (0
split)
##       num_total_rooms < 2.5       to the left,  agree=0.781, adj=0.125, (0
split)
##       price_persqft   < 0.7280547 to the right, agree=0.781, adj=0.125, (0
split)
##
## Node number 11: 22 observations
##   mean=528727.3, MSE=1.257356e+10
##
## Node number 12: 19 observations
##   mean=367844.7, MSE=8.995931e+09
##
## Node number 13: 39 observations
```

```
##   mean=536500, MSE=7.59891e+09
##
## Node number 14: 16 observations
##   mean=596000, MSE=1.695512e+10
##
## Node number 15: 10 observations
##   mean=774400, MSE=1.068544e+10
##
## Node number 18: 106 observations
##   mean=255325.5, MSE=6.709224e+09
##
## Node number 19: 12 observations
##   mean=391416.7, MSE=4.925576e+09
##
## Node number 20: 8 observations
##   mean=236943.8, MSE=5.856097e+09
##
## Node number 21: 24 observations
##   mean=439983.3, MSE=8.987341e+09
```

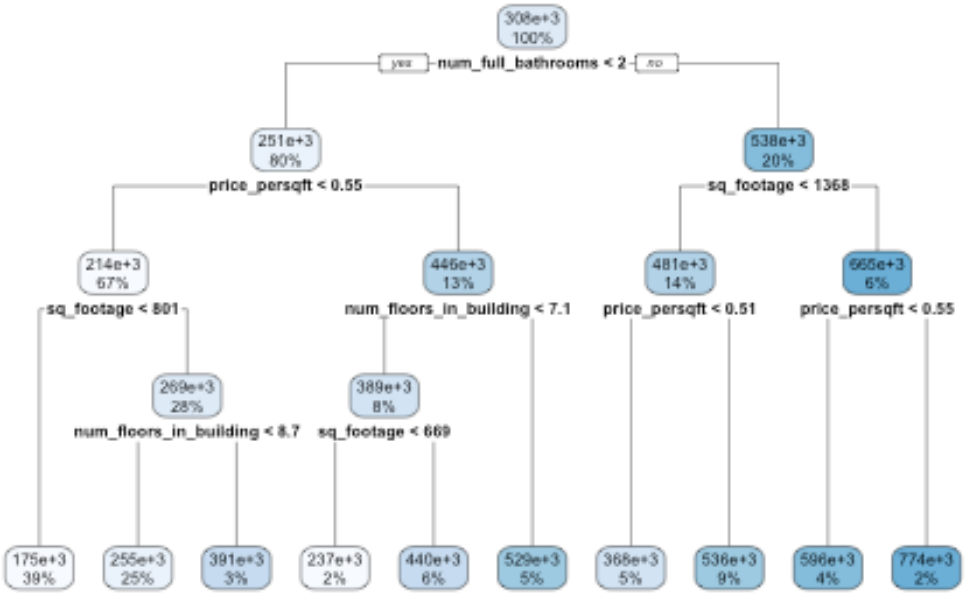
```
e3 <- yhat3 - Ytest
sqrt(sum(e3^2)/106)
```

```
## [1] 100728.2
```

```
mod3$cptable
```

```
##            CP nsplit rel error    xerror       xstd
## 1  0.42055632      0 1.0000000 1.0060020 0.07564038
## 2  0.18550984      1 0.5794437 0.6916319 0.06685520
## 3  0.04663919      2 0.3939338 0.4535066 0.04200201
## 4  0.04582955      3 0.3472947 0.4130265 0.03798872
## 5  0.02759198      4 0.3014651 0.3662018 0.03528808
## 6  0.01926389      5 0.2738731 0.3389620 0.03424167
## 7  0.01878043      6 0.2546092 0.3295965 0.03443944
## 8  0.01515856      7 0.2358288 0.3207510 0.03215055
## 9  0.01487063      8 0.2206703 0.3148164 0.03203113
## 10 0.01000000      9 0.2057996 0.2889764 0.02949261
```

```
rpart.plot(mod3)
```



Define function to get optimal cp and minimum error

```
get_cp <- function(x) {
  min    <- which.min(x$cptable[, "xerror"])
  cp <- x$cptable[min, "CP"]
}
get_min_error <- function(x) {
  min    <- which.min(x$cptable[, "xerror"])
  xerror <- x$cptable[min, "xerror"]
}
```

Get optimal tree

```
optimal_tree <- rpart(
  formula = Ytrain ~ .,
  data    = Xtrain,
  method  = "anova",
  control = list(minsplit = 11, maxdepth = 8, cp = 0.01)
)
summary(optimal_tree)

## Call:
## rpart(formula = Ytrain ~ ., data = Xtrain, method = "anova",
##     control = list(minsplit = 11, maxdepth = 8, cp = 0.01))
```

```
##   n= 422
##
##            CP nsplit rel error    xerror       xstd
## 1  0.42055632      0 1.0000000 1.0077405 0.07598962
## 2  0.18550984      1 0.5794437 0.6340607 0.05885281
## 3  0.04663919      2 0.3939338 0.4089786 0.03779693
## 4  0.04582955      3 0.3472947 0.4039400 0.03817255
## 5  0.02759198      4 0.3014651 0.3654430 0.03655211
## 6  0.01926389      5 0.2738731 0.3609597 0.03650736
## 7  0.01878043      6 0.2546092 0.3614732 0.03634978
## 8  0.01515856      7 0.2358288 0.3483807 0.03418525
## 9  0.01487063      8 0.2206703 0.3480160 0.03423637
## 10 0.01000000      9 0.2057996 0.3491763 0.03545829
##
## Variable importance
##     num_full_bathrooms             sq_footage           monthly_cost
##                     19                     17                     11
##          price_persqft            total_taxes        num_total_rooms
##                     10                      9                      8
##             coop_condo      approx_year_built     num_half_bathrooms
##                      7                      7                      4
##           num_bedrooms num_floors_in_building      pct_tax_deductibl
##                      3                      3                      2
##        parking_charges
##                      1
##
## Node number 1: 422 observations,    complexity param=0.4205563
##   mean=308191.7, MSE=3.121006e+10
##   left son=2 (338 obs) right son=3 (84 obs)
##   Primary splits:
##       num_full_bathrooms < 1.5       to the left,  improve=0.4205563, (0
missing)
##       price_persqft      < 0.5527845 to the left,  improve=0.3922844, (0
missing)
##       coop_condo         splits as  LR, improve=0.3754617, (0 missing)
##       approx_year_built  < 1970.5    to the left,  improve=0.3463094, (0
missing)
##       sq_footage         < 969.225   to the left,  improve=0.3139306, (0
missing)
##   Surrogate splits:
##       sq_footage         < 1133.495  to the left,  agree=0.917, adj=0.583,
(0 split)
##       total_taxes        < 4391.615  to the left,  agree=0.870, adj=0.345,
(0 split)
##       num_total_rooms    < 5.5       to the left,  agree=0.860, adj=0.298,
(0 split)
##       monthly_cost       < 1274.5    to the left,  agree=0.844, adj=0.214,
(0 split)
##       num_half_bathrooms < 0.455     to the right, agree=0.839, adj=0.190,
(0 split)
```

```
##
## Node number 2: 338 observations,    complexity param=0.1855098
##   mean=251077.9, MSE=1.663718e+10
##   left son=4 (284 obs) right son=5 (54 obs)
##   Primary splits:
##       price_persqft     < 0.5541679 to the left,  improve=0.4344879, (0
missing)
##       coop_condo        splits as  LR, improve=0.3876538, (0 missing)
##       approx_year_built < 1970.5    to the left,  improve=0.3067765, (0
missing)
##       sq_footage        < 857.955   to the left,  improve=0.1567627, (0
missing)
##       total_taxes       < 4198.69   to the left,  improve=0.1488824, (0
missing)
##   Surrogate splits:
##       coop_condo        splits as  LR, agree=0.956, adj=0.722, (0 split)
##       approx_year_built < 1970.5    to the left,  agree=0.944, adj=0.648,
(0 split)
##       monthly_cost      < 390.5     to the right, agree=0.902, adj=0.389,
(0 split)
##       pct_tax_deductibl < 34.945    to the right, agree=0.864, adj=0.148,
(0 split)
##       total_taxes       < 4463.51   to the left,  agree=0.861, adj=0.130,
(0 split)
##
## Node number 3: 84 observations,    complexity param=0.04582955
##   mean=538006.5, MSE=2.390811e+10
##   left son=6 (58 obs) right son=7 (26 obs)
##   Primary splits:
##       sq_footage             < 1368.09   to the left,  improve=0.3005579,
(0 missing)
##       price_persqft          < 0.4700775 to the left,  improve=0.2564064,
(0 missing)
##       num_floors_in_building < 14.5      to the left,  improve=0.2206020,
(0 missing)
##       parking_charges        < 69.735    to the left,  improve=0.2094910,
(0 missing)
##       total_taxes            < 4976.01   to the left,  improve=0.2043087,
(0 missing)
##   Surrogate splits:
##       monthly_cost           < 1461.5    to the left,  agree=0.857,
adj=0.538, (0 split)
##       num_bedrooms           < 2.5       to the left,  agree=0.833,
adj=0.462, (0 split)
##       num_floors_in_building < 11.525    to the left,  agree=0.798,
adj=0.346, (0 split)
##       num_total_rooms        < 6.5       to the left,  agree=0.774,
adj=0.269, (0 split)
##       total_taxes            < 4787.5    to the left,  agree=0.762,
adj=0.231, (0 split)
```

```
##
## Node number 4: 284 observations,    complexity param=0.04663919
##   mean=214004.2, MSE=7.534685e+09
##   left son=8 (166 obs) right son=9 (118 obs)
##   Primary splits:
##       sq_footage      < 800.5     to the left,  improve=0.2870613, (0
missing)
##       monthly_cost    < 761.5     to the left,  improve=0.2363109, (0
missing)
##       num_bedrooms    < 1.5       to the left,  improve=0.2009774, (0
missing)
##       num_total_rooms < 4.5       to the left,  improve=0.1218869, (0
missing)
##       price_persqft   < 0.5121664 to the left,  improve=0.1065049, (0
missing)
##   Surrogate splits:
##       num_bedrooms     < 1.5       to the left,  agree=0.891, adj=0.737,
(0 split)
##       num_total_rooms  < 3.5       to the left,  agree=0.835, adj=0.602,
(0 split)
##       monthly_cost     < 761.5     to the left,  agree=0.803, adj=0.525,
(0 split)
##       total_taxes      < 2831.22   to the left,  agree=0.704, adj=0.288,
(0 split)
##       dining_room_type splits as  LLR-R, agree=0.648, adj=0.153, (0 split)
##
## Node number 5: 54 observations,    complexity param=0.01926389
##   mean=446058.3, MSE=1.926355e+10
##   left son=10 (32 obs) right son=11 (22 obs)
##   Primary splits:
##       num_floors_in_building < 7.125     to the left,  improve=0.2439052,
(0 missing)
##       monthly_cost           < 304       to the left,  improve=0.2437085,
(0 missing)
##       total_taxes            < 2503.925  to the left,  improve=0.2306287,
(0 missing)
##       num_half_bathrooms     < 1.015     to the right, improve=0.2085708,
(0 missing)
##       sq_footage             < 669       to the left,  improve=0.2063172,
(0 missing)
##   Surrogate splits:
##       monthly_cost       < 401.5     to the left,  agree=0.796, adj=0.500,
(0 split)
##       num_half_bathrooms < 1.005     to the right, agree=0.741, adj=0.364,
(0 split)
##       pct_tax_deductibl  < 47.865    to the right, agree=0.722, adj=0.318,
(0 split)
##       total_taxes        < 4255.025  to the left,  agree=0.722, adj=0.318,
(0 split)
##       parking_charges    < 135.86    to the right, agree=0.685, adj=0.227,
```

```
(0 split)
##
## Node number 6: 58 observations,    complexity param=0.02759198
##   mean=481250.9, MSE=1.432215e+10
##   left son=12 (19 obs) right son=13 (39 obs)
##   Primary splits:
##       price_persqft     < 0.5087043 to the left,  improve=0.4374757, (0
missing)
##       approx_year_built < 1966.5    to the left,  improve=0.4026342, (0
missing)
##       coop_condo        splits as  LR, improve=0.3440129, (0 missing)
##       monthly_cost      < 812       to the right, improve=0.3440129, (0
missing)
##       parking_charges   < 73.5725   to the left,  improve=0.3059630, (0
missing)
##   Surrogate splits:
##       coop_condo        splits as  LR, agree=0.948, adj=0.842, (0 split)
##       monthly_cost      < 812       to the right, agree=0.948, adj=0.842,
(0 split)
##       approx_year_built < 1963.5    to the left,  agree=0.914, adj=0.737,
(0 split)
##       sq_footage        < 1143.5    to the right, agree=0.862, adj=0.579,
(0 split)
##       parking_charges   < 73.5725   to the left,  agree=0.793, adj=0.368,
(0 split)
##
## Node number 7: 26 observations,    complexity param=0.01487063
##   mean=664615.4, MSE=2.207662e+10
##   left son=14 (16 obs) right son=15 (10 obs)
##   Primary splits:
##       price_persqft          < 0.5451458 to the left,  improve=0.3412168,
(0 missing)
##       total_taxes            < 5018.9    to the left,  improve=0.3039461,
(0 missing)
##       parking_charges        < 141.8825  to the left,  improve=0.2528833,
(0 missing)
##       community_district_num < 27.5      to the left,  improve=0.2358061,
(0 missing)
##       num_half_bathrooms     < 0.945     to the right, improve=0.1572672,
(0 missing)
##   Surrogate splits:
##       total_taxes            < 5018.9    to the left,  agree=0.808,
adj=0.5, (0 split)
##       num_floors_in_building < 21        to the left,  agree=0.769,
adj=0.4, (0 split)
##       approx_year_built      < 1970.5    to the left,  agree=0.731,
adj=0.3, (0 split)
##       pct_tax_deductibl      < 35.69     to the right, agree=0.731,
adj=0.3, (0 split)
##       monthly_cost           < 2211      to the left,  agree=0.731,
```

```
adj=0.3, (0 split)
##
## Node number 8: 166 observations
##   mean=174793.3, MSE=3.347287e+09
##
## Node number 9: 118 observations,    complexity param=0.01515856
##   mean=269165.3, MSE=8.219769e+09
##   left son=18 (106 obs) right son=19 (12 obs)
##   Primary splits:
##       num_floors_in_building < 8.705     to the left,  improve=0.2058369,
(0 missing)
##       parking_charges        < 88.07     to the left,  improve=0.1792023,
(0 missing)
##       price_persqft          < 0.4516464 to the left,  improve=0.1596936,
(0 missing)
##       walk_score             < 91.5      to the left,  improve=0.1334127,
(0 missing)
##       monthly_cost           < 1048      to the left,  improve=0.1099628,
(0 missing)
##   Surrogate splits:
##       parking_charges   < 395.855   to the left,  agree=0.915, adj=0.167,
(0 split)
##       approx_year_built < 1964.5    to the left,  agree=0.907, adj=0.083,
(0 split)
##
## Node number 10: 32 observations,    complexity param=0.01878043
##   mean=389223.4, MSE=1.593423e+10
##   left son=20 (8 obs) right son=21 (24 obs)
##   Primary splits:
##       sq_footage      < 669       to the left,  improve=0.4851004, (0
missing)
##       num_total_rooms < 3.5       to the left,  improve=0.2499396, (0
missing)
##       parking_charges < 143.74    to the right, improve=0.2106907, (0
missing)
##       monthly_cost    < 297       to the left,  improve=0.1928689, (0
missing)
##       total_taxes     < 2503.925  to the left,  improve=0.1837103, (0
missing)
##   Surrogate splits:
##       num_bedrooms    < 0.5       to the left,  agree=0.812, adj=0.250, (0
split)
##       monthly_cost    < 177       to the left,  agree=0.812, adj=0.250, (0
split)
##       num_total_rooms < 2.5       to the left,  agree=0.781, adj=0.125, (0
split)
##       price_persqft   < 0.7280547 to the right, agree=0.781, adj=0.125, (0
split)
##
## Node number 11: 22 observations
```

```
##   mean=528727.3, MSE=1.257356e+10
##
## Node number 12: 19 observations
##   mean=367844.7, MSE=8.995931e+09
##
## Node number 13: 39 observations
##   mean=536500, MSE=7.59891e+09
##
## Node number 14: 16 observations
##   mean=596000, MSE=1.695512e+10
##
## Node number 15: 10 observations
##   mean=774400, MSE=1.068544e+10
##
## Node number 18: 106 observations
##   mean=255325.5, MSE=6.709224e+09
##
## Node number 19: 12 observations
##   mean=391416.7, MSE=4.925576e+09
##
## Node number 20: 8 observations
##   mean=236943.8, MSE=5.856097e+09
##
## Node number 21: 24 observations
##   mean=439983.3, MSE=8.987341e+09
```

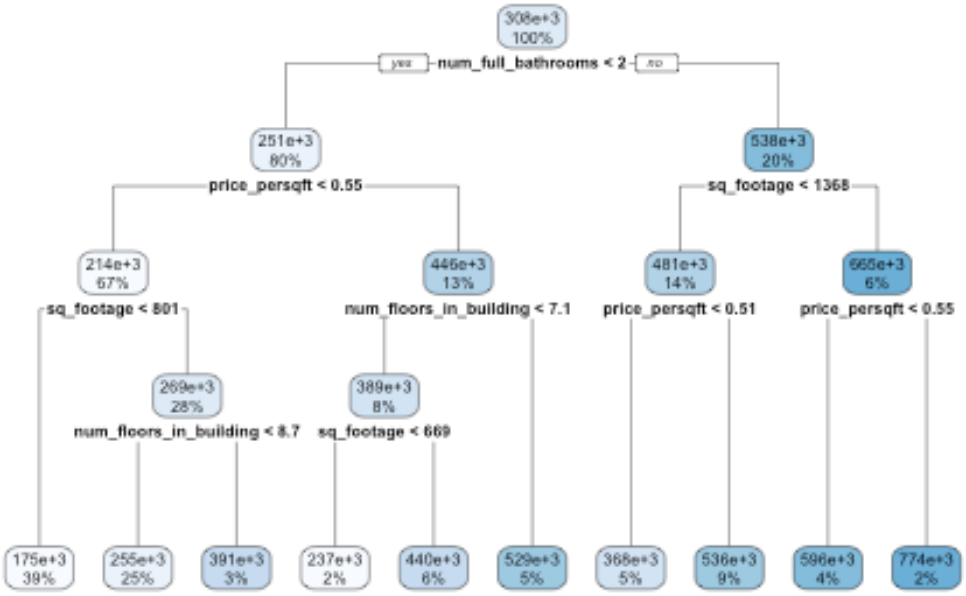
```
pred <- predict(optimal_tree, newdata = Xtrain)
RMSE(pred = pred, obs = Ytrain)
```

```
## [1] 80143.74
```

```
Tss = RMSE(pred = Ytrain, obs = mean(Ytrain))
 1-RMSE(pred = pred, obs = Ytrain)/Tss
```

```
## [1] 0.5463486
```

```
rpart.plot(optimal_tree)
```



Random forest

```
r_f1 <- randomForest(
  formula = Ytrain ~ .,
  data    = Xtrain
)
r_f1
```

```
##
## Call:
##  randomForest(formula = Ytrain ~ ., data = Xtrain)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 9
##
##           Mean of squared residuals: 5801632044
##                     % Var explained: 81.41
```

```
# print min mse index
which.min(r_f1$mse)
```

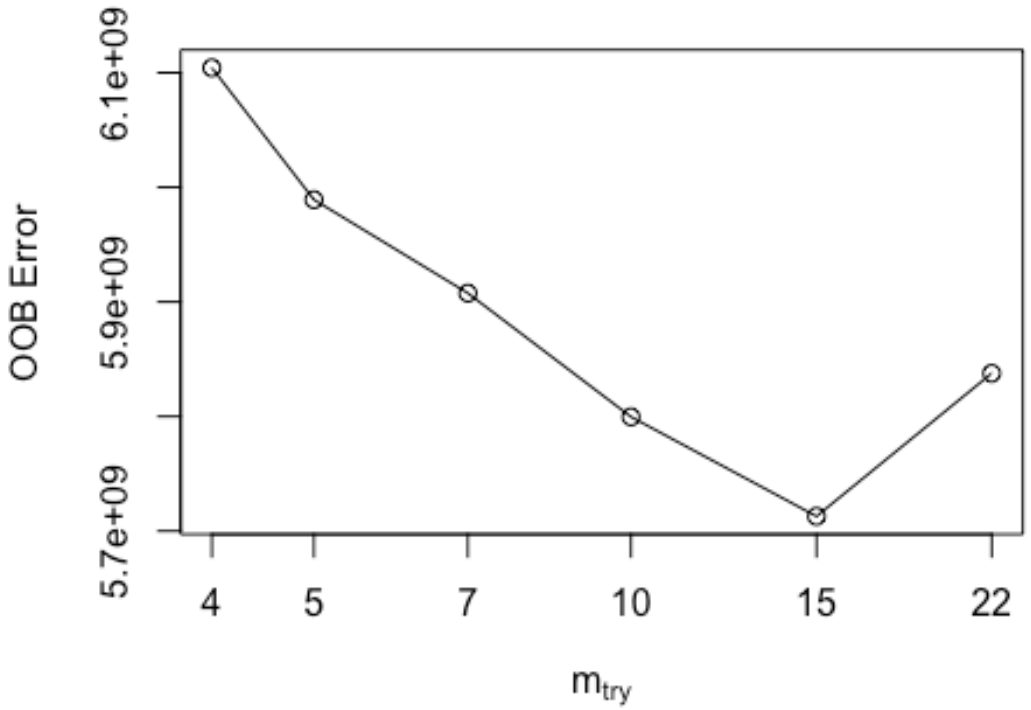
```
## [1] 166
```

```r
# RMSE of this optimal random forest
sqrt(r_f1$mse[which.min(r_f1$mse)])

## [1] 75776.4

features <- setdiff(names(Xtrain), Ytrain)
set.seed(1988)
r_f2 <- tuneRF(
  x          = Xtrain,
  y          = Ytrain,
  ntreeTry   = 500,
  mtryStart  = 5,
  stepFactor = 1.5,
  improve    = 0.01,
  trace      = FALSE
)

## -0.01921603 0.01
## 0.01362148 0.01
## 0.01827471 0.01
## 0.01496496 0.01
## -0.02184174 0.01
```



(UC Business Analytics R Programming Guide, n.d.)

Works Cited

UC Business Analytics R Programming Guide. (n.d.). Retrieved from https://uc-r.github.io/regression_trees

yiu, T. (n.d.). *Understanding Random Forest*. Retrieved from https://towardsdatascience.com/understanding-random-forest-58381e0602d2

End