Appendix 2 - House price prediction from 0.5 million record data set

### Step 1 : Collecting Data

The data set is collected in CSV format from kaggele and below is the reference for the same.

A.Sleem.(2018)Housepricing.[Online]. Available:<https://www.kaggle.com/greenwing1985/housepricing>

### Step 2 : Exploring, preprocessing and cleaning the data

Primary setup

knitr::opts\_knit$set(root.dir = '/Users/sobil/Documents/MSC/Sem 1/Data Mining & Machine Learning/Project/L5\_house/')  
remove(list = ls())  
set.seed(1)  
options(scipen=1)

loading alll the libraries required

library(data.table)  
library(boot)  
library(fpp2)

## Loading required package: ggplot2

## Loading required package: forecast

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

## Loading required package: fma

## Loading required package: expsmooth

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:boot':  
##   
## logit

library(leaps)

#### 1) reading the raw csv file

house <- fread("HousePrices\_HalfMil.csv")

#### 2) exploratory analysis

Structure of the house data frame

str(house)

## Classes 'data.table' and 'data.frame': 500000 obs. of 16 variables:  
## $ Area : int 164 84 190 75 148 124 58 249 243 242 ...  
## $ Garage : int 2 2 2 2 1 3 1 2 1 1 ...  
## $ FirePlace : int 0 0 4 4 4 3 0 1 0 2 ...  
## $ Baths : int 2 4 4 4 2 3 2 1 2 4 ...  
## $ White Marble : int 0 0 1 0 1 0 0 1 0 0 ...  
## $ Black Marble : int 1 0 0 0 0 1 0 0 0 0 ...  
## $ Indian Marble: int 0 1 0 1 0 0 1 0 1 1 ...  
## $ Floors : int 0 1 0 1 1 1 0 1 1 0 ...  
## $ City : int 3 2 2 1 2 1 3 1 1 2 ...  
## $ Solar : int 1 0 0 1 1 0 0 0 0 1 ...  
## $ Electric : int 1 0 0 1 0 0 1 1 0 0 ...  
## $ Fiber : int 1 0 1 1 0 1 1 0 0 0 ...  
## $ Glass Doors : int 1 1 0 1 1 1 1 1 0 0 ...  
## $ Swiming Pool : int 0 1 0 1 1 1 0 1 1 1 ...  
## $ Garden : int 0 1 0 1 1 1 1 0 0 0 ...  
## $ Prices : int 43800 37550 49500 50075 52400 54300 34400 50425 29575 22300 ...  
## - attr(\*, ".internal.selfref")=<externalptr>

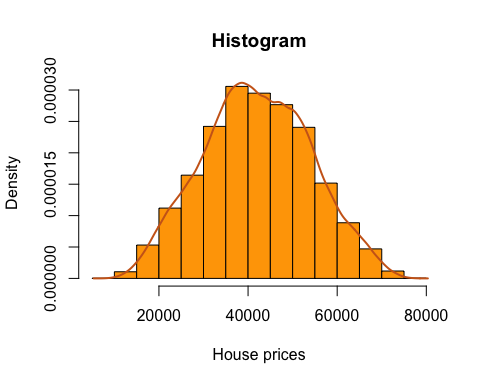
Summary of the house data frame

summary(house)

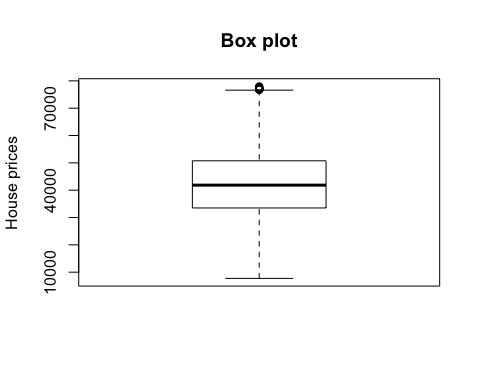
## Area Garage FirePlace Baths   
## Min. : 1.0 Min. :1.000 Min. :0.000 Min. :1.000   
## 1st Qu.: 63.0 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000   
## Median :125.0 Median :2.000 Median :2.000 Median :3.000   
## Mean :124.9 Mean :2.001 Mean :2.003 Mean :2.998   
## 3rd Qu.:187.0 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:4.000   
## Max. :249.0 Max. :3.000 Max. :4.000 Max. :5.000   
## White Marble Black Marble Indian Marble Floors   
## Min. :0.000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :0.000 Median :0.0000 Median :0.0000 Median :0.0000   
## Mean :0.333 Mean :0.3327 Mean :0.3343 Mean :0.4994   
## 3rd Qu.:1.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## City Solar Electric Fiber   
## Min. :1.000 Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:1.000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000   
## Median :2.000 Median :0.0000 Median :1.0000 Median :1.0000   
## Mean :2.001 Mean :0.4987 Mean :0.5007 Mean :0.5005   
## 3rd Qu.:3.000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :3.000 Max. :1.0000 Max. :1.0000 Max. :1.0000   
## Glass Doors Swiming Pool Garden Prices   
## Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. : 7725   
## 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:33500   
## Median :0.0000 Median :1.0000 Median :1.0000 Median :41850   
## Mean :0.4999 Mean :0.5004 Mean :0.5016 Mean :42050   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:50750   
## Max. :1.0000 Max. :1.0000 Max. :1.0000 Max. :77975

#### 3) check normal distribution of Prices

hist(house$Prices,  
 col = "orange",  
 border = "black",  
 prob = TRUE,  
 xlab = "House prices",  
 main = "Histogram")  
lines(density(house$Prices),  
 lwd = 2,  
 col = "chocolate3")



house.boxplot <- boxplot(house$Prices, main = "Box plot", ylab = "House prices")

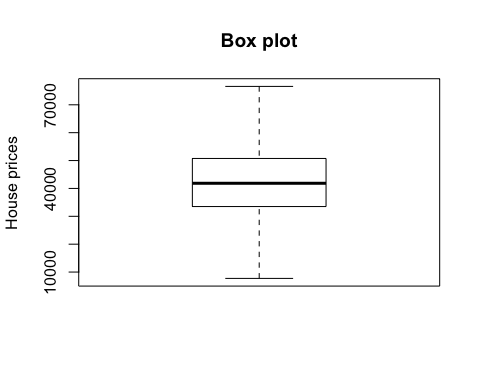


#### 4) removing outliers to make more genarlised model

house.boxplot$out

## [1] 76975 77225 77000 77175 77375 77525 76825 77700 76950 77075 77250 77975  
## [13] 76750 77225 76775 76800

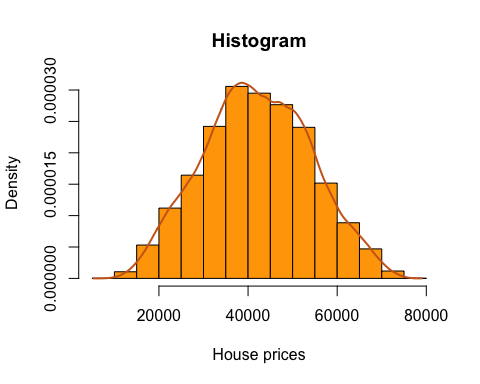
house <- subset(house, ! house$Prices %in% house.boxplot$out)  
house.boxplot <- boxplot(house$Prices, main = "Box plot", ylab = "House prices")



house.boxplot$out

## numeric(0)

hist(house$Prices,  
 col = "orange",  
 border = "black",  
 prob = TRUE,  
 xlab = "House prices",  
 main = "Histogram")  
lines(density(house$Prices),  
 lwd = 2,  
 col = "chocolate3")



house.boxplot$out

## numeric(0)

re-numbering the rows names

row.names(house) <- 1:499984

### Step 3 - Data transformation

cheking correlation

cor(house)

## Area Garage FirePlace Baths  
## Area 1.00000000000 -0.0009561749 3.171512e-04 -0.0004593863  
## Garage -0.00095617488 1.0000000000 1.260204e-03 -0.0036970899  
## FirePlace 0.00031715120 0.0012602041 1.000000e+00 0.0005954250  
## Baths -0.00045938627 -0.0036970899 5.954250e-04 1.0000000000  
## White Marble 0.00245686201 0.0004860091 8.982235e-04 0.0024346689  
## Black Marble -0.00144316109 0.0018742155 -8.954114e-04 -0.0027102665  
## Indian Marble -0.00101296442 -0.0023574470 -2.991059e-06 0.0002747674  
## Floors -0.00082493326 -0.0009699422 1.472555e-04 -0.0009210395  
## City -0.00351443467 0.0007451225 -2.573800e-04 -0.0009081267  
## Solar 0.00050591175 0.0014655403 -3.219511e-04 -0.0007673861  
## Electric -0.00016954633 0.0007447016 1.309616e-03 0.0010115692  
## Fiber 0.00006559031 -0.0006011681 1.780089e-03 -0.0007284917  
## Glass Doors -0.00127853023 -0.0022101528 -4.038464e-04 -0.0017088395  
## Swiming Pool 0.00061805058 0.0010929364 1.299237e-03 0.0022159658  
## Garden 0.00142202932 -0.0006733884 2.326840e-04 0.0017158806  
## Prices 0.14760173932 0.1001963924 8.904238e-02 0.1449912494  
## White Marble Black Marble Indian Marble Floors  
## Area 0.0024568620 -0.00144316109 -1.012964e-03 -0.00082493326  
## Garage 0.0004860091 0.00187421552 -2.357447e-03 -0.00096994219  
## FirePlace 0.0008982235 -0.00089541137 -2.991059e-06 0.00014725553  
## Baths 0.0024346689 -0.00271026646 2.747674e-04 -0.00092103948  
## White Marble 1.0000000000 -0.49888101866 -5.007113e-01 0.00003298180  
## Black Marble -0.4988810187 1.00000000000 -5.004070e-01 -0.00034552061  
## Indian Marble -0.5007113337 -0.50040700678 1.000000e+00 0.00031215124  
## Floors 0.0000329818 -0.00034552061 3.121512e-04 1.00000000000  
## City -0.0009160093 -0.00029667802 1.211395e-03 -0.00067982507  
## Solar -0.0016005727 0.00177222179 -1.711149e-04 -0.00266340286  
## Electric 0.0005187391 -0.00052889174 1.003497e-05 0.00005360457  
## Fiber -0.0006217479 0.00009601436 5.252191e-04 0.00134118361  
## Glass Doors -0.0004472523 0.00010855764 3.383727e-04 -0.00004036926  
## Swiming Pool -0.0018921204 0.00071005734 1.181007e-03 -0.00020690307  
## Garden 0.0009531764 0.00013600455 -1.088047e-03 -0.00049587593  
## Prices 0.4480976230 -0.07799497376 -3.697426e-01 0.61945121764  
## City Solar Electric Fiber  
## Area -0.0035144347 0.0005059118 -0.00016954633 0.00006559031  
## Garage 0.0007451225 0.0014655403 0.00074470160 -0.00060116814  
## FirePlace -0.0002573800 -0.0003219511 0.00130961628 0.00178008869  
## Baths -0.0009081267 -0.0007673861 0.00101156924 -0.00072849172  
## White Marble -0.0009160093 -0.0016005727 0.00051873907 -0.00062174794  
## Black Marble -0.0002966780 0.0017722218 -0.00052889174 0.00009601436  
## Indian Marble 0.0012113955 -0.0001711149 0.00001003497 0.00052521914  
## Floors -0.0006798251 -0.0026634029 0.00005360457 0.00134118361  
## City 1.0000000000 0.0004734082 0.00075332543 -0.00275531766  
## Solar 0.0004734082 1.0000000000 0.00187540582 0.00022638029  
## Electric 0.0007533254 0.0018754058 1.00000000000 -0.00033716112  
## Fiber -0.0027553177 0.0002263803 -0.00033716112 1.00000000000  
## Glass Doors 0.0007305603 -0.0008287956 0.00106040627 -0.00229981066  
## Swiming Pool 0.0003273864 -0.0004577177 0.00057090464 0.00413134355  
## Garden 0.0012019720 -0.0042715465 0.00077184703 -0.00002697347  
## Prices 0.2331825181 0.0083951719 0.05236927271 0.48460589251  
## Glass Doors Swiming Pool Garden Prices  
## Area -0.00127853023 0.0006180506 0.00142202932 0.147601739  
## Garage -0.00221015284 0.0010929364 -0.00067338843 0.100196392  
## FirePlace -0.00040384643 0.0012992366 0.00023268400 0.089042379  
## Baths -0.00170883950 0.0022159658 0.00171588060 0.144991249  
## White Marble -0.00044725227 -0.0018921204 0.00095317639 0.448097623  
## Black Marble 0.00010855764 0.0007100573 0.00013600455 -0.077994974  
## Indian Marble 0.00033837269 0.0011810067 -0.00108804715 -0.369742607  
## Floors -0.00004036926 -0.0002069031 -0.00049587593 0.619451218  
## City 0.00073056032 0.0003273864 0.00120197205 0.233182518  
## Solar -0.00082879564 -0.0004577177 -0.00427154650 0.008395172  
## Electric 0.00106040627 0.0005709046 0.00077184703 0.052369273  
## Fiber -0.00229981066 0.0041313435 -0.00002697347 0.484605893  
## Glass Doors 1.00000000000 0.0004002688 0.00332508465 0.181907887  
## Swiming Pool 0.00040026878 1.0000000000 -0.00019888790 0.001798530  
## Garden 0.00332508465 -0.0001988879 1.00000000000 0.001529416  
## Prices 0.18190788732 0.0017985297 0.00152941620 1.000000000

making a new df with only correlated columns

chouse <- house  
chouse <- chouse[,-c(2,3,6,10,11,14,15)]  
  
chouse.colnames <-colnames(chouse)  
chouse.colnames[3] <- "White\_Marbel"  
chouse.colnames[4] <- "Indian\_Marbel"  
colnames(chouse) <- chouse.colnames

checking correlation

cor(chouse)

## Area Baths White\_Marbel Indian\_Marbel  
## Area 1.00000000000 -0.0004593863 0.0024568620 -0.0010129644  
## Baths -0.00045938627 1.0000000000 0.0024346689 0.0002747674  
## White\_Marbel 0.00245686201 0.0024346689 1.0000000000 -0.5007113337  
## Indian\_Marbel -0.00101296442 0.0002747674 -0.5007113337 1.0000000000  
## Floors -0.00082493326 -0.0009210395 0.0000329818 0.0003121512  
## City -0.00351443467 -0.0009081267 -0.0009160093 0.0012113955  
## Fiber 0.00006559031 -0.0007284917 -0.0006217479 0.0005252191  
## Glass Doors -0.00127853023 -0.0017088395 -0.0004472523 0.0003383727  
## Prices 0.14760173932 0.1449912494 0.4480976230 -0.3697426073  
## Floors City Fiber Glass Doors  
## Area -0.00082493326 -0.0035144347 0.00006559031 -0.00127853023  
## Baths -0.00092103948 -0.0009081267 -0.00072849172 -0.00170883950  
## White\_Marbel 0.00003298180 -0.0009160093 -0.00062174794 -0.00044725227  
## Indian\_Marbel 0.00031215124 0.0012113955 0.00052521914 0.00033837269  
## Floors 1.00000000000 -0.0006798251 0.00134118361 -0.00004036926  
## City -0.00067982507 1.0000000000 -0.00275531766 0.00073056032  
## Fiber 0.00134118361 -0.0027553177 1.00000000000 -0.00229981066  
## Glass Doors -0.00004036926 0.0007305603 -0.00229981066 1.00000000000  
## Prices 0.61945121764 0.2331825181 0.48460589251 0.18190788732  
## Prices  
## Area 0.1476017  
## Baths 0.1449912  
## White\_Marbel 0.4480976  
## Indian\_Marbel -0.3697426  
## Floors 0.6194512  
## City 0.2331825  
## Fiber 0.4846059  
## Glass Doors 0.1819079  
## Prices 1.0000000

### Step 4 : Training a model on the data

model 1

house.fit1 <- glm(Prices ~ . + White\_Marbel:Indian\_Marbel , data = chouse)

### Step 5 : Evaluating the model

Cheking the summary of the model

summary(house.fit1)

##   
## Call:  
## glm(formula = Prices ~ . + White\_Marbel:Indian\_Marbel, data = chouse)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3769.7 -1254.9 -4.7 1245.7 3761.8   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11265.35304 10.91311 1032.3 <2e-16 \*\*\*  
## Area 24.98769 0.03434 727.6 <2e-16 \*\*\*  
## Baths 1247.61897 1.74349 715.6 <2e-16 \*\*\*  
## White\_Marbel 9000.64343 6.04412 1489.2 <2e-16 \*\*\*  
## Indian\_Marbel -5005.84224 6.03795 -829.1 <2e-16 \*\*\*  
## Floors 14997.32899 4.93130 3041.3 <2e-16 \*\*\*  
## City 3501.43505 3.02091 1159.1 <2e-16 \*\*\*  
## Fiber 11751.93391 4.93132 2383.1 <2e-16 \*\*\*  
## `Glass Doors` 4444.83567 4.93131 901.4 <2e-16 \*\*\*  
## White\_Marbel:Indian\_Marbel NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 3039597)  
##   
## Null deviance: 7.3309e+13 on 499983 degrees of freedom  
## Residual deviance: 1.5197e+12 on 499975 degrees of freedom  
## AIC: 8882283  
##   
## Number of Fisher Scoring iterations: 2

### Step 6 : Improving the model

#### 1) Adding all the possible relation to the model

model 2 adding the correlation independent variabels

house.fit2 <- update(house.fit1, ~ . - White\_Marbel:Indian\_Marbel)  
summary(house.fit2) #8882283

##   
## Call:  
## glm(formula = Prices ~ Area + Baths + White\_Marbel + Indian\_Marbel +   
## Floors + City + Fiber + `Glass Doors`, data = chouse)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3769.7 -1254.9 -4.7 1245.7 3761.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11265.35304 10.91311 1032.3 <2e-16 \*\*\*  
## Area 24.98769 0.03434 727.6 <2e-16 \*\*\*  
## Baths 1247.61897 1.74349 715.6 <2e-16 \*\*\*  
## White\_Marbel 9000.64343 6.04412 1489.2 <2e-16 \*\*\*  
## Indian\_Marbel -5005.84224 6.03795 -829.1 <2e-16 \*\*\*  
## Floors 14997.32899 4.93130 3041.3 <2e-16 \*\*\*  
## City 3501.43505 3.02091 1159.1 <2e-16 \*\*\*  
## Fiber 11751.93391 4.93132 2383.1 <2e-16 \*\*\*  
## `Glass Doors` 4444.83567 4.93131 901.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 3039597)  
##   
## Null deviance: 7.3309e+13 on 499983 degrees of freedom  
## Residual deviance: 1.5197e+12 on 499975 degrees of freedom  
## AIC: 8882283  
##   
## Number of Fisher Scoring iterations: 2

accuracy(house.fit2)

## ME RMSE MAE MPE MAPE MASE  
## Training set 5.791451e-09 1743.428 1427.536 -0.2387077 3.763029 0.1442308

house.k10.fit2.err <- cv.glm(data = chouse,glmfit = house.fit2, K = 10)  
house.k10.fit2.err$delta# 3039673

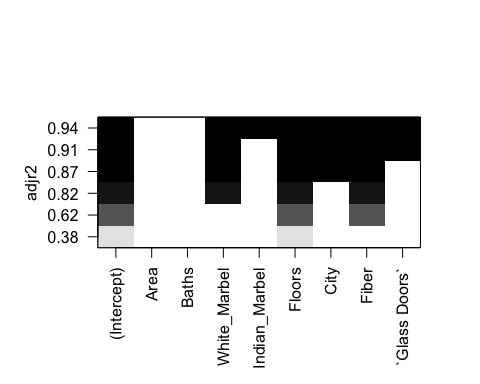
## [1] 3039655 3039649

#checking r2  
house.fit2.lm <- lm(Prices ~ ., data = chouse)  
summary(house.fit2.lm) # 0.9793

##   
## Call:  
## lm(formula = Prices ~ ., data = chouse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3769.7 -1254.9 -4.7 1245.7 3761.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11265.35304 10.91311 1032.3 <2e-16 \*\*\*  
## Area 24.98769 0.03434 727.6 <2e-16 \*\*\*  
## Baths 1247.61897 1.74349 715.6 <2e-16 \*\*\*  
## White\_Marbel 9000.64343 6.04412 1489.2 <2e-16 \*\*\*  
## Indian\_Marbel -5005.84224 6.03795 -829.1 <2e-16 \*\*\*  
## Floors 14997.32899 4.93130 3041.3 <2e-16 \*\*\*  
## City 3501.43505 3.02091 1159.1 <2e-16 \*\*\*  
## Fiber 11751.93391 4.93132 2383.1 <2e-16 \*\*\*  
## `Glass Doors` 4444.83567 4.93131 901.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1743 on 499975 degrees of freedom  
## Multiple R-squared: 0.9793, Adjusted R-squared: 0.9793   
## F-statistic: 2.952e+06 on 8 and 499975 DF, p-value: < 2.2e-16

#### 2) check the best fit

house.fit2.bestFit1 <- regsubsets(Prices ~ Area + Baths + White\_Marbel + Indian\_Marbel +   
 Floors + City + Fiber + `Glass Doors`, data = chouse, nbest = 1, nvmax = 6)  
par(mfrow = c(1,1))  
#subsets(house.fit2.bestFit1, statistic = "adjr2", max.size = 6, min.size = 1)  
plot(house.fit2.bestFit1, scale = "adjr2")



4 predictors

house.fit4 <- glm(Prices ~ White\_Marbel + Floors + City + Fiber, data = chouse)  
summary(house.fit4) #9783831

##   
## Call:  
## glm(formula = Prices ~ White\_Marbel + Floors + City + Fiber,   
## data = chouse)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -14061.6 -2997.2 -4.2 3009.2 14020.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17865.211 18.748 952.9 <2e-16 \*\*\*  
## White\_Marbel 11526.007 12.889 894.3 <2e-16 \*\*\*  
## Floors 14989.394 12.148 1233.9 <2e-16 \*\*\*  
## City 3491.541 7.442 469.2 <2e-16 \*\*\*  
## Fiber 11738.342 12.148 966.3 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 18446727)  
##   
## Null deviance: 7.3309e+13 on 499983 degrees of freedom  
## Residual deviance: 9.2230e+12 on 499979 degrees of freedom  
## AIC: 9783831  
##   
## Number of Fisher Scoring iterations: 2

house.k10.fit4.err <- cv.glm(data = chouse,glmfit = house.fit4, K = 10)  
house.k10.fit4.err$delta# 18446855

## [1] 18446975 18446952

#checking r2  
house.fit4.lm <- lm(Prices ~ White\_Marbel + Floors + City + Fiber, data = chouse)  
summary(house.fit4.lm) # 0.8742

##   
## Call:  
## lm(formula = Prices ~ White\_Marbel + Floors + City + Fiber, data = chouse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14061.6 -2997.2 -4.2 3009.2 14020.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17865.211 18.748 952.9 <2e-16 \*\*\*  
## White\_Marbel 11526.007 12.889 894.3 <2e-16 \*\*\*  
## Floors 14989.394 12.148 1233.9 <2e-16 \*\*\*  
## City 3491.541 7.442 469.2 <2e-16 \*\*\*  
## Fiber 11738.342 12.148 966.3 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4295 on 499979 degrees of freedom  
## Multiple R-squared: 0.8742, Adjusted R-squared: 0.8742   
## F-statistic: 8.685e+05 on 4 and 499979 DF, p-value: < 2.2e-16

5 predictors

house.fit5 <- glm(Prices ~ White\_Marbel + Floors + City + Fiber + `Glass Doors`, data = chouse)  
summary(house.fit5) #9628934

##   
## Call:  
## glm(formula = Prices ~ White\_Marbel + Floors + City + Fiber +   
## `Glass Doors`, data = chouse)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -11849.7 -2573.4 7.6 2569.7 11835.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15647.064 16.880 927.0 <2e-16 \*\*\*  
## White\_Marbel 11528.115 11.039 1044.3 <2e-16 \*\*\*  
## Floors 14989.557 10.405 1440.6 <2e-16 \*\*\*  
## City 3489.575 6.374 547.5 <2e-16 \*\*\*  
## Fiber 11748.531 10.405 1129.1 <2e-16 \*\*\*  
## `Glass Doors` 4433.692 10.405 426.1 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 13532329)  
##   
## Null deviance: 7.3309e+13 on 499983 degrees of freedom  
## Residual deviance: 6.7659e+12 on 499978 degrees of freedom  
## AIC: 9628934  
##   
## Number of Fisher Scoring iterations: 2

house.k10.fit5.err <- cv.glm(data = chouse,glmfit = house.fit5, K = 10)  
house.k10.fit5.err$delta# 13532564

## [1] 13532487 13532470

#checking r2  
house.fit5.lm <- lm(Prices ~ White\_Marbel + Floors + City + Fiber + `Glass Doors`, data = chouse)  
summary(house.fit5.lm) # 0.9077

##   
## Call:  
## lm(formula = Prices ~ White\_Marbel + Floors + City + Fiber +   
## `Glass Doors`, data = chouse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11849.7 -2573.4 7.6 2569.7 11835.7   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 15647.064 16.880 927.0 <2e-16 \*\*\*  
## White\_Marbel 11528.115 11.039 1044.3 <2e-16 \*\*\*  
## Floors 14989.557 10.405 1440.6 <2e-16 \*\*\*  
## City 3489.575 6.374 547.5 <2e-16 \*\*\*  
## Fiber 11748.531 10.405 1129.1 <2e-16 \*\*\*  
## `Glass Doors` 4433.692 10.405 426.1 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3679 on 499978 degrees of freedom  
## Multiple R-squared: 0.9077, Adjusted R-squared: 0.9077   
## F-statistic: 9.835e+05 on 5 and 499978 DF, p-value: < 2.2e-16

6 predictors

house.fit6 <- glm(Prices ~ White\_Marbel + Indian\_Marbel + Floors + City + Fiber + `Glass Doors`, data = chouse)  
summary(house.fit6) #9445048

##   
## Call:  
## glm(formula = Prices ~ White\_Marbel + Indian\_Marbel + Floors +   
## City + Fiber + `Glass Doors`, data = chouse)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -9370.3 -2152.8 0.8 2147.3 9351.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 18145.883 15.012 1208.8 <2e-16 \*\*\*  
## White\_Marbel 9023.390 10.611 850.4 <2e-16 \*\*\*  
## Indian\_Marbel -4997.261 10.600 -471.4 <2e-16 \*\*\*  
## Floors 14991.108 8.657 1731.6 <2e-16 \*\*\*  
## City 3491.752 5.303 658.4 <2e-16 \*\*\*  
## Fiber 11749.548 8.657 1357.2 <2e-16 \*\*\*  
## `Glass Doors` 4434.231 8.657 512.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 9367959)  
##   
## Null deviance: 7.3309e+13 on 499983 degrees of freedom  
## Residual deviance: 4.6838e+12 on 499977 degrees of freedom  
## AIC: 9445048  
##   
## Number of Fisher Scoring iterations: 2

house.k10.fit6.err <- cv.glm(data = chouse,glmfit = house.fit6, K = 10)  
house.k10.fit6.err$delta# 9368131

## [1] 9368083 9368069

#checking r2  
house.fit6.lm <- lm(Prices ~ White\_Marbel + Indian\_Marbel + Floors + City + Fiber + `Glass Doors`, data = chouse)  
summary(house.fit6.lm) # 0.9361

##   
## Call:  
## lm(formula = Prices ~ White\_Marbel + Indian\_Marbel + Floors +   
## City + Fiber + `Glass Doors`, data = chouse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9370.3 -2152.8 0.8 2147.3 9351.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 18145.883 15.012 1208.8 <2e-16 \*\*\*  
## White\_Marbel 9023.390 10.611 850.4 <2e-16 \*\*\*  
## Indian\_Marbel -4997.261 10.600 -471.4 <2e-16 \*\*\*  
## Floors 14991.108 8.657 1731.6 <2e-16 \*\*\*  
## City 3491.752 5.303 658.4 <2e-16 \*\*\*  
## Fiber 11749.548 8.657 1357.2 <2e-16 \*\*\*  
## `Glass Doors` 4434.231 8.657 512.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3061 on 499977 degrees of freedom  
## Multiple R-squared: 0.9361, Adjusted R-squared: 0.9361   
## F-statistic: 1.221e+06 on 6 and 499977 DF, p-value: < 2.2e-16

#### 3) Best model performance is house.fit2

par(mfrow = c(2,2))  
summary(house.fit2)

##   
## Call:  
## glm(formula = Prices ~ Area + Baths + White\_Marbel + Indian\_Marbel +   
## Floors + City + Fiber + `Glass Doors`, data = chouse)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3769.7 -1254.9 -4.7 1245.7 3761.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11265.35304 10.91311 1032.3 <2e-16 \*\*\*  
## Area 24.98769 0.03434 727.6 <2e-16 \*\*\*  
## Baths 1247.61897 1.74349 715.6 <2e-16 \*\*\*  
## White\_Marbel 9000.64343 6.04412 1489.2 <2e-16 \*\*\*  
## Indian\_Marbel -5005.84224 6.03795 -829.1 <2e-16 \*\*\*  
## Floors 14997.32899 4.93130 3041.3 <2e-16 \*\*\*  
## City 3501.43505 3.02091 1159.1 <2e-16 \*\*\*  
## Fiber 11751.93391 4.93132 2383.1 <2e-16 \*\*\*  
## `Glass Doors` 4444.83567 4.93131 901.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 3039597)  
##   
## Null deviance: 7.3309e+13 on 499983 degrees of freedom  
## Residual deviance: 1.5197e+12 on 499975 degrees of freedom  
## AIC: 8882283  
##   
## Number of Fisher Scoring iterations: 2

summary(house.fit2.lm)

##   
## Call:  
## lm(formula = Prices ~ ., data = chouse)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3769.7 -1254.9 -4.7 1245.7 3761.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11265.35304 10.91311 1032.3 <2e-16 \*\*\*  
## Area 24.98769 0.03434 727.6 <2e-16 \*\*\*  
## Baths 1247.61897 1.74349 715.6 <2e-16 \*\*\*  
## White\_Marbel 9000.64343 6.04412 1489.2 <2e-16 \*\*\*  
## Indian\_Marbel -5005.84224 6.03795 -829.1 <2e-16 \*\*\*  
## Floors 14997.32899 4.93130 3041.3 <2e-16 \*\*\*  
## City 3501.43505 3.02091 1159.1 <2e-16 \*\*\*  
## Fiber 11751.93391 4.93132 2383.1 <2e-16 \*\*\*  
## `Glass Doors` 4444.83567 4.93131 901.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1743 on 499975 degrees of freedom  
## Multiple R-squared: 0.9793, Adjusted R-squared: 0.9793   
## F-statistic: 2.952e+06 on 8 and 499975 DF, p-value: < 2.2e-16

plot(house.fit2)

