library(RCurl)

## Loading required package: bitops

library(MASS)  
library(leaps)

houseData <- read.csv("file:///G:/Ryerson-BigData/capstone-R/CKME-136/data/kc\_house\_data.csv")  
houseData$date<-NULL  
houseData$id<-NULL  
colnames(houseData)

## [1] "price" "bedrooms" "bathrooms" "sqft\_living"   
## [5] "sqft\_lot" "floors" "waterfront" "view"   
## [9] "condition" "grade" "sqft\_above" "sqft\_basement"  
## [13] "yr\_built" "yr\_renovated" "zipcode" "lat"   
## [17] "long" "sqft\_living15" "sqft\_lot15"

## Let us now use the forward selection algorithm using stepAIC.

full = lm(price~.,data=houseData)  
null = lm(price~1,data=houseData)  
stepF = stepAIC(null, scope=list(lower=null, upper=full), direction='forward', trace=TRUE)

## Start: AIC=553875.8  
## price ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + sqft\_living 1 1.4356e+15 1.4773e+15 539204  
## + grade 1 1.2976e+15 1.6153e+15 541134  
## + sqft\_above 1 1.0682e+15 1.8447e+15 544004  
## + sqft\_living15 1 9.9816e+14 1.9148e+15 544810  
## + bathrooms 1 8.0329e+14 2.1096e+15 546904  
## + view 1 4.5978e+14 2.4531e+15 550165  
## + sqft\_basement 1 3.0544e+14 2.6075e+15 551484  
## + bedrooms 1 2.7696e+14 2.6360e+15 551718  
## + lat 1 2.7455e+14 2.6384e+15 551738  
## + waterfront 1 2.0668e+14 2.7062e+15 552287  
## + floors 1 1.9209e+14 2.7208e+15 552403  
## + yr\_renovated 1 4.6564e+13 2.8664e+15 553529  
## + sqft\_lot 1 2.3417e+13 2.8895e+15 553703  
## + sqft\_lot15 1 1.9801e+13 2.8931e+15 553730  
## + yr\_built 1 8.4977e+12 2.9044e+15 553815  
## + zipcode 1 8.2451e+12 2.9047e+15 553817  
## + condition 1 3.8514e+12 2.9091e+15 553849  
## + long 1 1.3624e+12 2.9116e+15 553868  
## <none> 2.9129e+15 553876  
##   
## Step: AIC=539203.5  
## price ~ sqft\_living  
##   
## Df Sum of Sq RSS AIC  
## + lat 1 2.1314e+14 1.2641e+15 535838  
## + view 1 1.2362e+14 1.3537e+15 537317  
## + grade 1 1.2132e+14 1.3560e+15 537353  
## + waterfront 1 1.1024e+14 1.3670e+15 537529  
## + yr\_built 1 9.2854e+13 1.3844e+15 537802  
## + long 1 6.6817e+13 1.4105e+15 538205  
## + bedrooms 1 4.0635e+13 1.4366e+15 538603  
## + zipcode 1 2.2858e+13 1.4544e+15 538868  
## + yr\_renovated 1 2.2405e+13 1.4549e+15 538875  
## + sqft\_living15 1 2.0109e+13 1.4572e+15 538909  
## + condition 1 1.7605e+13 1.4597e+15 538946  
## + sqft\_lot15 1 6.4407e+12 1.4708e+15 539111  
## + sqft\_lot 1 3.0113e+12 1.4743e+15 539161  
## + sqft\_above 1 1.2165e+12 1.4761e+15 539188  
## + sqft\_basement 1 1.2165e+12 1.4761e+15 539188  
## + floors 1 2.2991e+11 1.4770e+15 539202  
## + bathrooms 1 1.4719e+11 1.4771e+15 539203  
## <none> 1.4773e+15 539204  
##   
## Step: AIC=535838  
## price ~ sqft\_living + lat  
##   
## Df Sum of Sq RSS AIC  
## + view 1 1.2663e+14 1.1375e+15 533559  
## + waterfront 1 1.1646e+14 1.1477e+15 533751  
## + grade 1 8.8423e+13 1.1757e+15 534273  
## + yr\_built 1 5.1904e+13 1.2122e+15 534934  
## + long 1 3.6167e+13 1.2280e+15 535213  
## + bedrooms 1 3.2254e+13 1.2319e+15 535281  
## + condition 1 1.9095e+13 1.2450e+15 535511  
## + yr\_renovated 1 1.8897e+13 1.2452e+15 535515  
## + sqft\_living15 1 1.8325e+13 1.2458e+15 535524  
## + sqft\_lot15 1 1.2429e+12 1.2629e+15 535819  
## + zipcode 1 4.4621e+11 1.2637e+15 535832  
## <none> 1.2641e+15 535838  
## + sqft\_lot 1 1.0913e+11 1.2640e+15 535838  
## + sqft\_above 1 1.0387e+11 1.2640e+15 535838  
## + sqft\_basement 1 1.0387e+11 1.2640e+15 535838  
## + bathrooms 1 2.2942e+09 1.2641e+15 535840  
## + floors 1 2.9322e+07 1.2641e+15 535840  
##   
## Step: AIC=533558.7  
## price ~ sqft\_living + lat + view  
##   
## Df Sum of Sq RSS AIC  
## + grade 1 7.7085e+13 1.0604e+15 532044  
## + waterfront 1 4.8301e+13 1.0892e+15 532623  
## + yr\_built 1 2.9685e+13 1.1078e+15 532989  
## + bedrooms 1 2.0105e+13 1.1174e+15 533175  
## + long 1 1.8126e+13 1.1194e+15 533214  
## + condition 1 1.3259e+13 1.1242e+15 533307  
## + yr\_renovated 1 1.1033e+13 1.1265e+15 533350  
## + sqft\_living15 1 9.7773e+12 1.1277e+15 533374  
## + sqft\_above 1 5.6493e+12 1.1319e+15 533453  
## + sqft\_basement 1 5.6493e+12 1.1319e+15 533453  
## + sqft\_lot15 1 1.8222e+12 1.1357e+15 533526  
## + zipcode 1 1.3136e+12 1.1362e+15 533536  
## + floors 1 7.9084e+11 1.1367e+15 533546  
## + sqft\_lot 1 3.9207e+11 1.1371e+15 533553  
## + bathrooms 1 1.9270e+11 1.1373e+15 533557  
## <none> 1.1375e+15 533559  
##   
## Step: AIC=532044.1  
## price ~ sqft\_living + lat + view + grade  
##   
## Df Sum of Sq RSS AIC  
## + yr\_built 1 8.9146e+13 9.7128e+14 530148  
## + waterfront 1 5.0218e+13 1.0102e+15 530998  
## + condition 1 2.5997e+13 1.0344e+15 531510  
## + long 1 2.2309e+13 1.0381e+15 531587  
## + yr\_renovated 1 1.4312e+13 1.0461e+15 531752  
## + bedrooms 1 1.0398e+13 1.0500e+15 531833  
## + floors 1 3.9309e+12 1.0565e+15 531966  
## + bathrooms 1 2.2781e+12 1.0581e+15 532000  
## + sqft\_lot15 1 1.3272e+12 1.0591e+15 532019  
## + sqft\_lot 1 2.0910e+11 1.0602e+15 532042  
## + sqft\_above 1 1.3720e+11 1.0603e+15 532043  
## + sqft\_basement 1 1.3720e+11 1.0603e+15 532043  
## + sqft\_living15 1 1.1809e+11 1.0603e+15 532044  
## <none> 1.0604e+15 532044  
## + zipcode 1 7.8101e+10 1.0603e+15 532045  
##   
## Step: AIC=530148.3  
## price ~ sqft\_living + lat + view + grade + yr\_built  
##   
## Df Sum of Sq RSS AIC  
## + waterfront 1 5.0449e+13 9.2083e+14 528997  
## + bedrooms 1 1.1098e+13 9.6018e+14 529902  
## + zipcode 1 6.4623e+12 9.6481e+14 530006  
## + bathrooms 1 5.2656e+12 9.6601e+14 530033  
## + condition 1 4.2739e+12 9.6700e+14 530055  
## + long 1 2.8391e+12 9.6844e+14 530087  
## + yr\_renovated 1 2.3436e+12 9.6893e+14 530098  
## + floors 1 2.1809e+12 9.6910e+14 530102  
## + sqft\_above 1 2.1769e+12 9.6910e+14 530102  
## + sqft\_basement 1 2.1769e+12 9.6910e+14 530102  
## + sqft\_lot15 1 1.1384e+12 9.7014e+14 530125  
## + sqft\_living15 1 6.4656e+11 9.7063e+14 530136  
## + sqft\_lot 1 2.8898e+11 9.7099e+14 530144  
## <none> 9.7128e+14 530148  
##   
## Step: AIC=528997.4  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront  
##   
## Df Sum of Sq RSS AIC  
## + bedrooms 1 9.0057e+12 9.1182e+14 528787  
## + zipcode 1 6.3395e+12 9.1449e+14 528850  
## + bathrooms 1 5.4031e+12 9.1543e+14 528872  
## + condition 1 4.4331e+12 9.1639e+14 528895  
## + long 1 2.5647e+12 9.1826e+14 528939  
## + floors 1 1.6628e+12 9.1917e+14 528960  
## + sqft\_above 1 1.4511e+12 9.1938e+14 528965  
## + sqft\_basement 1 1.4511e+12 9.1938e+14 528965  
## + yr\_renovated 1 1.2489e+12 9.1958e+14 528970  
## + sqft\_lot15 1 1.1644e+12 9.1966e+14 528972  
## + sqft\_living15 1 9.9743e+11 9.1983e+14 528976  
## + sqft\_lot 1 2.2143e+11 9.2061e+14 528994  
## <none> 9.2083e+14 528997  
##   
## Step: AIC=528787  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms  
##   
## Df Sum of Sq RSS AIC  
## + bathrooms 1 9.0102e+12 9.0281e+14 528574  
## + zipcode 1 6.9168e+12 9.0491e+14 528624  
## + condition 1 5.2627e+12 9.0656e+14 528664  
## + long 1 2.8458e+12 9.0898e+14 528721  
## + sqft\_lot15 1 1.9499e+12 9.0987e+14 528743  
## + floors 1 1.7197e+12 9.1010e+14 528748  
## + yr\_renovated 1 1.1626e+12 9.1066e+14 528761  
## + sqft\_above 1 1.1004e+12 9.1072e+14 528763  
## + sqft\_basement 1 1.1004e+12 9.1072e+14 528763  
## + sqft\_living15 1 8.3834e+11 9.1098e+14 528769  
## + sqft\_lot 1 5.6135e+11 9.1126e+14 528776  
## <none> 9.1182e+14 528787  
##   
## Step: AIC=528574.4  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms  
##   
## Df Sum of Sq RSS AIC  
## + zipcode 1 7.7118e+12 8.9510e+14 528391  
## + condition 1 4.8311e+12 8.9798e+14 528460  
## + long 1 2.0291e+12 9.0078e+14 528528  
## + sqft\_above 1 1.6201e+12 9.0119e+14 528538  
## + sqft\_basement 1 1.6201e+12 9.0119e+14 528538  
## + sqft\_living15 1 1.4832e+12 9.0133e+14 528541  
## + sqft\_lot15 1 1.4206e+12 9.0139e+14 528542  
## + yr\_renovated 1 4.2334e+11 9.0239e+14 528566  
## + floors 1 3.9100e+11 9.0242e+14 528567  
## + sqft\_lot 1 3.6234e+11 9.0245e+14 528568  
## <none> 9.0281e+14 528574  
##   
## Step: AIC=528391  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms + zipcode  
##   
## Df Sum of Sq RSS AIC  
## + long 1 9.6955e+12 8.8540e+14 528158  
## + condition 1 3.5280e+12 8.9157e+14 528308  
## + sqft\_lot15 1 2.1877e+12 8.9291e+14 528340  
## + sqft\_above 1 1.1628e+12 8.9394e+14 528365  
## + sqft\_basement 1 1.1628e+12 8.9394e+14 528365  
## + floors 1 1.0460e+12 8.9405e+14 528368  
## + sqft\_lot 1 7.2300e+11 8.9438e+14 528376  
## + sqft\_living15 1 4.3262e+11 8.9467e+14 528383  
## + yr\_renovated 1 3.8775e+11 8.9471e+14 528384  
## <none> 8.9510e+14 528391  
##   
## Step: AIC=528157.6  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms + zipcode + long  
##   
## Df Sum of Sq RSS AIC  
## + condition 1 3.2700e+12 8.8213e+14 528080  
## + sqft\_above 1 2.8185e+12 8.8259e+14 528091  
## + sqft\_basement 1 2.8185e+12 8.8259e+14 528091  
## + sqft\_living15 1 1.5701e+12 8.8383e+14 528121  
## + floors 1 8.8103e+11 8.8452e+14 528138  
## + sqft\_lot15 1 7.8011e+11 8.8462e+14 528141  
## + yr\_renovated 1 5.1267e+11 8.8489e+14 528147  
## <none> 8.8540e+14 528158  
## + sqft\_lot 1 8.0929e+10 8.8532e+14 528158  
##   
## Step: AIC=528079.6  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms + zipcode + long + condition  
##   
## Df Sum of Sq RSS AIC  
## + sqft\_above 1 3.6952e+12 8.7844e+14 527991  
## + sqft\_basement 1 3.6952e+12 8.7844e+14 527991  
## + sqft\_living15 1 1.7368e+12 8.8040e+14 528039  
## + floors 1 1.3208e+12 8.8081e+14 528049  
## + yr\_renovated 1 1.1069e+12 8.8103e+14 528055  
## + sqft\_lot15 1 7.9924e+11 8.8134e+14 528062  
## <none> 8.8213e+14 528080  
## + sqft\_lot 1 7.3301e+10 8.8206e+14 528080  
##   
## Step: AIC=527990.9  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms + zipcode + long + condition + sqft\_above  
##   
## Df Sum of Sq RSS AIC  
## + sqft\_living15 1 1.2623e+12 8.7718e+14 527962  
## + yr\_renovated 1 1.0677e+12 8.7737e+14 527967  
## + sqft\_lot15 1 8.9305e+11 8.7755e+14 527971  
## + floors 1 1.2603e+11 8.7831e+14 527990  
## + sqft\_lot 1 1.0887e+11 8.7833e+14 527990  
## <none> 8.7844e+14 527991  
##   
## Step: AIC=527961.8  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms + zipcode + long + condition + sqft\_above +   
## sqft\_living15  
##   
## Df Sum of Sq RSS AIC  
## + yr\_renovated 1 1.2135e+12 8.7596e+14 527934  
## + sqft\_lot15 1 9.2340e+11 8.7625e+14 527941  
## + floors 1 2.3644e+11 8.7694e+14 527958  
## + sqft\_lot 1 8.4109e+10 8.7709e+14 527962  
## <none> 8.7718e+14 527962  
##   
## Step: AIC=527933.9  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms + zipcode + long + condition + sqft\_above +   
## sqft\_living15 + yr\_renovated  
##   
## Df Sum of Sq RSS AIC  
## + sqft\_lot15 1 9.2931e+11 8.7503e+14 527913  
## + floors 1 1.8361e+11 8.7578e+14 527931  
## <none> 8.7596e+14 527934  
## + sqft\_lot 1 7.8513e+10 8.7589e+14 527934  
##   
## Step: AIC=527913  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms + zipcode + long + condition + sqft\_above +   
## sqft\_living15 + yr\_renovated + sqft\_lot15  
##   
## Df Sum of Sq RSS AIC  
## + sqft\_lot 1 2.8295e+11 8.7475e+14 527908  
## + floors 1 1.3147e+11 8.7490e+14 527912  
## <none> 8.7503e+14 527913  
##   
## Step: AIC=527908  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms + zipcode + long + condition + sqft\_above +   
## sqft\_living15 + yr\_renovated + sqft\_lot15 + sqft\_lot  
##   
## Df Sum of Sq RSS AIC  
## + floors 1 1.4017e+11 8.7461e+14 527907  
## <none> 8.7475e+14 527908  
##   
## Step: AIC=527906.5  
## price ~ sqft\_living + lat + view + grade + yr\_built + waterfront +   
## bedrooms + bathrooms + zipcode + long + condition + sqft\_above +   
## sqft\_living15 + yr\_renovated + sqft\_lot15 + sqft\_lot + floors  
##   
## Df Sum of Sq RSS AIC  
## <none> 8.7461e+14 527907

summary(stepF)

##   
## Call:  
## lm(formula = price ~ sqft\_living + lat + view + grade + yr\_built +   
## waterfront + bedrooms + bathrooms + zipcode + long + condition +   
## sqft\_above + sqft\_living15 + yr\_renovated + sqft\_lot15 +   
## sqft\_lot + floors, data = houseData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1291725 -99229 -9739 77583 4333222   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.690e+06 2.931e+06 2.282 0.02249 \*   
## sqft\_living 1.501e+02 4.385e+00 34.227 < 2e-16 \*\*\*  
## lat 6.027e+05 1.073e+04 56.149 < 2e-16 \*\*\*  
## view 5.287e+04 2.140e+03 24.705 < 2e-16 \*\*\*  
## grade 9.589e+04 2.153e+03 44.542 < 2e-16 \*\*\*  
## yr\_built -2.620e+03 7.266e+01 -36.062 < 2e-16 \*\*\*  
## waterfront 5.830e+05 1.736e+04 33.580 < 2e-16 \*\*\*  
## bedrooms -3.577e+04 1.892e+03 -18.906 < 2e-16 \*\*\*  
## bathrooms 4.114e+04 3.254e+03 12.645 < 2e-16 \*\*\*  
## zipcode -5.824e+02 3.299e+01 -17.657 < 2e-16 \*\*\*  
## long -2.147e+05 1.313e+04 -16.349 < 2e-16 \*\*\*  
## condition 2.639e+04 2.351e+03 11.221 < 2e-16 \*\*\*  
## sqft\_above 3.113e+01 4.360e+00 7.139 9.71e-13 \*\*\*  
## sqft\_living15 2.168e+01 3.448e+00 6.289 3.26e-10 \*\*\*  
## yr\_renovated 1.981e+01 3.656e+00 5.420 6.03e-08 \*\*\*  
## sqft\_lot15 -3.826e-01 7.327e-02 -5.222 1.78e-07 \*\*\*  
## sqft\_lot 1.286e-01 4.792e-02 2.683 0.00729 \*\*   
## floors 6.690e+03 3.596e+03 1.860 0.06285 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 201200 on 21595 degrees of freedom  
## Multiple R-squared: 0.6997, Adjusted R-squared: 0.6995   
## F-statistic: 2960 on 17 and 21595 DF, p-value: < 2.2e-16

subsets = regsubsets(price~.,data=houseData,nbest=1,)

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,  
## force.in = force.in, : 1 linear dependencies found

## Reordering variables and trying again:

sub.sum = summary(subsets)  
as.data.frame(sub.sum$outmat)

## bedrooms bathrooms sqft\_living sqft\_lot floors waterfront view  
## 1 ( 1 ) \*   
## 2 ( 1 ) \*   
## 3 ( 1 ) \* \*  
## 4 ( 1 ) \*   
## 5 ( 1 ) \* \*   
## 6 ( 1 ) \* \* \*  
## 7 ( 1 ) \* \* \* \*  
## 8 ( 1 ) \* \* \* \* \*  
## 9 ( 1 ) \* \* \* \*  
## condition grade sqft\_above sqft\_basement yr\_built yr\_renovated  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 ) \* \*   
## 5 ( 1 ) \* \*   
## 6 ( 1 ) \* \*   
## 7 ( 1 ) \* \*   
## 8 ( 1 ) \* \*   
## 9 ( 1 ) \* \*   
## zipcode lat long sqft\_living15 sqft\_lot15  
## 1 ( 1 )   
## 2 ( 1 ) \*   
## 3 ( 1 ) \*   
## 4 ( 1 ) \*   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \*   
## 9 ( 1 ) \* \* \*

## Let us now use the backward selection algorithm using stepAIC.

full = lm(price~.,data=houseData)  
null = lm(price~1,data=houseData)  
stepF = stepAIC(full, direction= 'backward', trace=TRUE)

## Start: AIC=527906.5  
## price ~ bedrooms + bathrooms + sqft\_living + sqft\_lot + floors +   
## waterfront + view + condition + grade + sqft\_above + sqft\_basement +   
## yr\_built + yr\_renovated + zipcode + lat + long + sqft\_living15 +   
## sqft\_lot15  
##   
##   
## Step: AIC=527906.5  
## price ~ bedrooms + bathrooms + sqft\_living + sqft\_lot + floors +   
## waterfront + view + condition + grade + sqft\_above + yr\_built +   
## yr\_renovated + zipcode + lat + long + sqft\_living15 + sqft\_lot15  
##   
## Df Sum of Sq RSS AIC  
## <none> 8.7461e+14 527907  
## - floors 1 1.4017e+11 8.7475e+14 527908  
## - sqft\_lot 1 2.9164e+11 8.7490e+14 527912  
## - sqft\_lot15 1 1.1046e+12 8.7572e+14 527932  
## - yr\_renovated 1 1.1897e+12 8.7580e+14 527934  
## - sqft\_living15 1 1.6017e+12 8.7621e+14 527944  
## - sqft\_above 1 2.0640e+12 8.7668e+14 527955  
## - condition 1 5.0994e+12 8.7971e+14 528030  
## - bathrooms 1 6.4764e+12 8.8109e+14 528064  
## - long 1 1.0826e+13 8.8544e+14 528170  
## - zipcode 1 1.2626e+13 8.8724e+14 528214  
## - bedrooms 1 1.4476e+13 8.8909e+14 528259  
## - view 1 2.4720e+13 8.9933e+14 528507  
## - waterfront 1 4.5671e+13 9.2028e+14 529005  
## - sqft\_living 1 4.7447e+13 9.2206e+14 529046  
## - yr\_built 1 5.2669e+13 9.2728e+14 529168  
## - grade 1 8.0354e+13 9.5497e+14 529804  
## - lat 1 1.2769e+14 1.0023e+15 530850

summary(stepF)

##   
## Call:  
## lm(formula = price ~ bedrooms + bathrooms + sqft\_living + sqft\_lot +   
## floors + waterfront + view + condition + grade + sqft\_above +   
## yr\_built + yr\_renovated + zipcode + lat + long + sqft\_living15 +   
## sqft\_lot15, data = houseData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1291725 -99229 -9739 77583 4333222   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.690e+06 2.931e+06 2.282 0.02249 \*   
## bedrooms -3.577e+04 1.892e+03 -18.906 < 2e-16 \*\*\*  
## bathrooms 4.114e+04 3.254e+03 12.645 < 2e-16 \*\*\*  
## sqft\_living 1.501e+02 4.385e+00 34.227 < 2e-16 \*\*\*  
## sqft\_lot 1.286e-01 4.792e-02 2.683 0.00729 \*\*   
## floors 6.690e+03 3.596e+03 1.860 0.06285 .   
## waterfront 5.830e+05 1.736e+04 33.580 < 2e-16 \*\*\*  
## view 5.287e+04 2.140e+03 24.705 < 2e-16 \*\*\*  
## condition 2.639e+04 2.351e+03 11.221 < 2e-16 \*\*\*  
## grade 9.589e+04 2.153e+03 44.542 < 2e-16 \*\*\*  
## sqft\_above 3.113e+01 4.360e+00 7.139 9.71e-13 \*\*\*  
## yr\_built -2.620e+03 7.266e+01 -36.062 < 2e-16 \*\*\*  
## yr\_renovated 1.981e+01 3.656e+00 5.420 6.03e-08 \*\*\*  
## zipcode -5.824e+02 3.299e+01 -17.657 < 2e-16 \*\*\*  
## lat 6.027e+05 1.073e+04 56.149 < 2e-16 \*\*\*  
## long -2.147e+05 1.313e+04 -16.349 < 2e-16 \*\*\*  
## sqft\_living15 2.168e+01 3.448e+00 6.289 3.26e-10 \*\*\*  
## sqft\_lot15 -3.826e-01 7.327e-02 -5.222 1.78e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 201200 on 21595 degrees of freedom  
## Multiple R-squared: 0.6997, Adjusted R-squared: 0.6995   
## F-statistic: 2960 on 17 and 21595 DF, p-value: < 2.2e-16

subsets = regsubsets(price~.,data=houseData,nbest=1,)

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,  
## force.in = force.in, : 1 linear dependencies found

## Reordering variables and trying again:

sub.sum = summary(subsets)  
as.data.frame(sub.sum$outmat)

## bedrooms bathrooms sqft\_living sqft\_lot floors waterfront view  
## 1 ( 1 ) \*   
## 2 ( 1 ) \*   
## 3 ( 1 ) \* \*  
## 4 ( 1 ) \*   
## 5 ( 1 ) \* \*   
## 6 ( 1 ) \* \* \*  
## 7 ( 1 ) \* \* \* \*  
## 8 ( 1 ) \* \* \* \* \*  
## 9 ( 1 ) \* \* \* \*  
## condition grade sqft\_above sqft\_basement yr\_built yr\_renovated  
## 1 ( 1 )   
## 2 ( 1 )   
## 3 ( 1 )   
## 4 ( 1 ) \* \*   
## 5 ( 1 ) \* \*   
## 6 ( 1 ) \* \*   
## 7 ( 1 ) \* \*   
## 8 ( 1 ) \* \*   
## 9 ( 1 ) \* \*   
## zipcode lat long sqft\_living15 sqft\_lot15  
## 1 ( 1 )   
## 2 ( 1 ) \*   
## 3 ( 1 ) \*   
## 4 ( 1 ) \*   
## 5 ( 1 ) \*   
## 6 ( 1 ) \*   
## 7 ( 1 ) \*   
## 8 ( 1 ) \*   
## 9 ( 1 ) \* \* \*

#since floor is having very less contribution to price.we are not including floor in our model.  
str(houseData)

## 'data.frame': 21613 obs. of 19 variables:  
## $ price : num 221900 538000 180000 604000 510000 ...  
## $ bedrooms : int 3 3 2 4 3 4 3 3 3 3 ...  
## $ bathrooms : num 1 2.25 1 3 2 4.5 2.25 1.5 1 2.5 ...  
## $ sqft\_living : int 1180 2570 770 1960 1680 5420 1715 1060 1780 1890 ...  
## $ sqft\_lot : int 5650 7242 10000 5000 8080 101930 6819 9711 7470 6560 ...  
## $ floors : num 1 2 1 1 1 1 2 1 1 2 ...  
## $ waterfront : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ view : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ condition : int 3 3 3 5 3 3 3 3 3 3 ...  
## $ grade : int 7 7 6 7 8 11 7 7 7 7 ...  
## $ sqft\_above : int 1180 2170 770 1050 1680 3890 1715 1060 1050 1890 ...  
## $ sqft\_basement: int 0 400 0 910 0 1530 0 0 730 0 ...  
## $ yr\_built : int 1955 1951 1933 1965 1987 2001 1995 1963 1960 2003 ...  
## $ yr\_renovated : int 0 1991 0 0 0 0 0 0 0 0 ...  
## $ zipcode : int 98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 ...  
## $ lat : num 47.5 47.7 47.7 47.5 47.6 ...  
## $ long : num -122 -122 -122 -122 -122 ...  
## $ sqft\_living15: int 1340 1690 2720 1360 1800 4760 2238 1650 1780 2390 ...  
## $ sqft\_lot15 : int 5650 7639 8062 5000 7503 101930 6819 9711 8113 7570 ...

sapply(houseData, is.numeric)

## price bedrooms bathrooms sqft\_living sqft\_lot   
## TRUE TRUE TRUE TRUE TRUE   
## floors waterfront view condition grade   
## TRUE TRUE TRUE TRUE TRUE   
## sqft\_above sqft\_basement yr\_built yr\_renovated zipcode   
## TRUE TRUE TRUE TRUE TRUE   
## lat long sqft\_living15 sqft\_lot15   
## TRUE TRUE TRUE TRUE

houseDatas <- houseData[ , sapply(houseData, is.numeric)]  
cor(houseDatas)

## price bedrooms bathrooms sqft\_living  
## price 1.00000000 0.308349598 0.52513751 0.70203505  
## bedrooms 0.30834960 1.000000000 0.51588364 0.57667069  
## bathrooms 0.52513751 0.515883638 1.00000000 0.75466528  
## sqft\_living 0.70203505 0.576670693 0.75466528 1.00000000  
## sqft\_lot 0.08966086 0.031703243 0.08773966 0.17282566  
## floors 0.25679389 0.175428935 0.50065317 0.35394929  
## waterfront 0.26636943 -0.006582479 0.06374363 0.10381782  
## view 0.39729349 0.079531852 0.18773702 0.28461119  
## condition 0.03636179 0.028472104 -0.12498193 -0.05875259  
## grade 0.66743426 0.356966725 0.66498253 0.76270448  
## sqft\_above 0.60556730 0.477600161 0.68534248 0.87659660  
## sqft\_basement 0.32381602 0.303093375 0.28377003 0.43504297  
## yr\_built 0.05401153 0.154178069 0.50601944 0.31804877  
## yr\_renovated 0.12643379 0.018840823 0.05073898 0.05536293  
## zipcode -0.05320285 -0.152668487 -0.20386627 -0.19943004  
## lat 0.30700348 -0.008931010 0.02457295 0.05252946  
## long 0.02162624 0.129472975 0.22304184 0.24022330  
## sqft\_living15 0.58537890 0.391637524 0.56863429 0.75642026  
## sqft\_lot15 0.08244715 0.029244224 0.08717536 0.18328555  
## sqft\_lot floors waterfront view  
## price 0.089660861 0.256793888 0.266369434 0.397293488  
## bedrooms 0.031703243 0.175428935 -0.006582479 0.079531852  
## bathrooms 0.087739662 0.500653173 0.063743629 0.187737024  
## sqft\_living 0.172825661 0.353949290 0.103817818 0.284611186  
## sqft\_lot 1.000000000 -0.005200991 0.021603683 0.074710106  
## floors -0.005200991 1.000000000 0.023698320 0.029443820  
## waterfront 0.021603683 0.023698320 1.000000000 0.401857351  
## view 0.074710106 0.029443820 0.401857351 1.000000000  
## condition -0.008958250 -0.263767946 0.016653157 0.045989737  
## grade 0.113621124 0.458182514 0.082774914 0.251320585  
## sqft\_above 0.183512281 0.523884710 0.072074592 0.167649344  
## sqft\_basement 0.015286202 -0.245704542 0.080587939 0.276946579  
## yr\_built 0.053080367 0.489319425 -0.026161086 -0.053439851  
## yr\_renovated 0.007643505 0.006338401 0.092884837 0.103917288  
## zipcode -0.129574486 -0.059120642 0.030284728 0.084826917  
## lat -0.085682788 0.049614131 -0.014273776 0.006156732  
## long 0.229520859 0.125419028 -0.041910200 -0.078399712  
## sqft\_living15 0.144608174 0.279885265 0.086463136 0.280439082  
## sqft\_lot15 0.718556752 -0.011269187 0.030703283 0.072574568  
## condition grade sqft\_above sqft\_basement  
## price 0.036361789 0.66743426 0.6055672984 0.32381602  
## bedrooms 0.028472104 0.35696673 0.4776001614 0.30309338  
## bathrooms -0.124981933 0.66498253 0.6853424759 0.28377003  
## sqft\_living -0.058752587 0.76270448 0.8765965987 0.43504297  
## sqft\_lot -0.008958250 0.11362112 0.1835122809 0.01528620  
## floors -0.263767946 0.45818251 0.5238847103 -0.24570454  
## waterfront 0.016653157 0.08277491 0.0720745917 0.08058794  
## view 0.045989737 0.25132058 0.1676493441 0.27694658  
## condition 1.000000000 -0.14467367 -0.1582136164 0.17410491  
## grade -0.144673671 1.00000000 0.7559229376 0.16839182  
## sqft\_above -0.158213616 0.75592294 1.0000000000 -0.05194331  
## sqft\_basement 0.174104914 0.16839182 -0.0519433068 1.00000000  
## yr\_built -0.361416562 0.44696320 0.4238983517 -0.13312410  
## yr\_renovated -0.060617787 0.01441428 0.0232846879 0.07132290  
## zipcode 0.003025524 -0.18486209 -0.2611899765 0.07484461  
## lat -0.014941006 0.11408406 -0.0008164986 0.11053796  
## long -0.106500448 0.19837215 0.3438030175 -0.14476477  
## sqft\_living15 -0.092824268 0.71320209 0.7318702924 0.20035498  
## sqft\_lot15 -0.003405523 0.11924790 0.1940498619 0.01727618  
## yr\_built yr\_renovated zipcode lat  
## price 0.05401153 0.126433793 -0.053202854 0.3070034800  
## bedrooms 0.15417807 0.018840823 -0.152668487 -0.0089310097  
## bathrooms 0.50601944 0.050738978 -0.203866274 0.0245729528  
## sqft\_living 0.31804877 0.055362927 -0.199430043 0.0525294622  
## sqft\_lot 0.05308037 0.007643505 -0.129574486 -0.0856827882  
## floors 0.48931942 0.006338401 -0.059120642 0.0496141310  
## waterfront -0.02616109 0.092884837 0.030284728 -0.0142737756  
## view -0.05343985 0.103917288 0.084826917 0.0061567321  
## condition -0.36141656 -0.060617787 0.003025524 -0.0149410064  
## grade 0.44696320 0.014414281 -0.184862093 0.1140840571  
## sqft\_above 0.42389835 0.023284688 -0.261189977 -0.0008164986  
## sqft\_basement -0.13312410 0.071322902 0.074844608 0.1105379580  
## yr\_built 1.00000000 -0.224873518 -0.346869178 -0.1481224021  
## yr\_renovated -0.22487352 1.000000000 0.064357057 0.0293976092  
## zipcode -0.34686918 0.064357057 1.000000000 0.2670479500  
## lat -0.14812240 0.029397609 0.267047950 1.0000000000  
## long 0.40935620 -0.068372369 -0.564071606 -0.1355117836  
## sqft\_living15 0.32622890 -0.002672555 -0.279032997 0.0488579321  
## sqft\_lot15 0.07095793 0.007853765 -0.147221069 -0.0864188072  
## long sqft\_living15 sqft\_lot15  
## price 0.02162624 0.585378904 0.082447153  
## bedrooms 0.12947298 0.391637524 0.029244224  
## bathrooms 0.22304184 0.568634290 0.087175361  
## sqft\_living 0.24022330 0.756420259 0.183285551  
## sqft\_lot 0.22952086 0.144608174 0.718556752  
## floors 0.12541903 0.279885265 -0.011269187  
## waterfront -0.04191020 0.086463136 0.030703283  
## view -0.07839971 0.280439082 0.072574568  
## condition -0.10650045 -0.092824268 -0.003405523  
## grade 0.19837215 0.713202093 0.119247897  
## sqft\_above 0.34380302 0.731870292 0.194049862  
## sqft\_basement -0.14476477 0.200354983 0.017276181  
## yr\_built 0.40935620 0.326228900 0.070957926  
## yr\_renovated -0.06837237 -0.002672555 0.007853765  
## zipcode -0.56407161 -0.279032997 -0.147221069  
## lat -0.13551178 0.048857932 -0.086418807  
## long 1.00000000 0.334604984 0.254451288  
## sqft\_living15 0.33460498 1.000000000 0.183191749  
## sqft\_lot15 0.25445129 0.183191749 1.000000000

## Using Step AIC we got below variables needs to part of model

## 1:sqft\_living

## 2:lat

## 3:grade

## 4:yr\_built

## 5:waterfront

## 6:bedrooms

## 7:bathrooms

## 8:zipcode

## 9:long

#since no any variable is highly correlated with each other  
houseData$floors<-NULL  
houseData$date<-NULL  
colnames(houseData)

## [1] "price" "bedrooms" "bathrooms" "sqft\_living"   
## [5] "sqft\_lot" "waterfront" "view" "condition"   
## [9] "grade" "sqft\_above" "sqft\_basement" "yr\_built"   
## [13] "yr\_renovated" "zipcode" "lat" "long"   
## [17] "sqft\_living15" "sqft\_lot15"

newhouseData <- subset(houseData, select = c(price,sqft\_living,lat,view,grade,yr\_built,waterfront,bedrooms,bathrooms,zipcode,long))  
  
set.seed(1)  
i=0.6  
storage <- list(c(), c(), c(),c())  
for(i in seq(from=0.6, to=0.9, by=0.01)){  
 rn\_train <- sample(nrow(newhouseData),floor(nrow(newhouseData)\*i))  
 train <- newhouseData[rn\_train,colnames(newhouseData)]  
 test <- newhouseData[-rn\_train,colnames(newhouseData)]  
 model<-lm(price~sqft\_living+lat+view+grade+yr\_built+waterfront+bedrooms+bathrooms+zipcode+long,data = train)  
 prediction <- predict(model,interval='prediction',newdata = test)  
 train\_prediction = fitted(model)  
 train\_rmse = sqrt(sum((train\_prediction-train$price)^2)/nrow(train))  
 test\_rmse = sqrt(sum((prediction - test$price)^2)/nrow(test))  
 storage[[1]]<-c(storage[[1]],i)  
 storage[[2]]<-c(storage[[2]],test\_rmse)  
 storage[[3]]<-c(storage[[3]],train\_rmse)  
   
}  
  
##find the LM with minimun training error  
RMSE = storage[[3]]  
minimumVal = min(RMSE)  
minimumVal

## [1] 196954.3

indx = which(RMSE==min(RMSE))  
indx

## [1] 14

storage[[1]][indx]

## [1] 0.73

cat("\nMinimum Training RMSE of Regression:",storage[[3]][indx],"\nRMSE of testing :",storage[[2]][indx], "\nTraining data Percentage:",storage[[1]][indx])

##   
## Minimum Training RMSE of Regression: 196954.3   
## RMSE of testing : 663350.1   
## Training data Percentage: 0.73

## Now we come to a conclusion that 73% Training data provides the Minimum RMSE

## 1: SET training Data = 73% &

## 2: Get model with coeeficient & Intercept

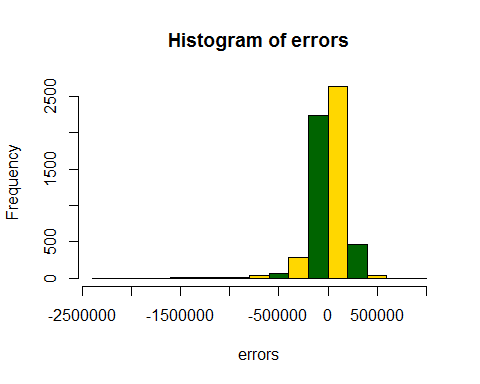
## 3: Draw the error of Histogram to get confidence with model

## 4: Find out how many data have less than 25% of error

set.seed(1)  
rn\_train <- sample(nrow(newhouseData),floor(nrow(newhouseData)\*storage[[1]][indx]))  
train <- newhouseData[rn\_train,colnames(newhouseData)]  
test <- newhouseData[-rn\_train,colnames(newhouseData)]  
modelXGen <- lm(price~sqft\_living+lat+view+grade+yr\_built+waterfront+bedrooms+bathrooms+zipcode+long,data = train)  
summary(modelXGen)

##   
## Call:  
## lm(formula = price ~ sqft\_living + lat + view + grade + yr\_built +   
## waterfront + bedrooms + bathrooms + zipcode + long, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1242635 -99486 -9406 78454 4345822   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.581e+07 3.267e+06 4.839 1.32e-06 \*\*\*  
## sqft\_living 1.755e+02 3.588e+00 48.918 < 2e-16 \*\*\*  
## lat 5.915e+05 1.258e+04 47.026 < 2e-16 \*\*\*  
## view 5.297e+04 2.470e+03 21.447 < 2e-16 \*\*\*  
## grade 1.039e+05 2.362e+03 43.983 < 2e-16 \*\*\*  
## yr\_built -2.806e+03 7.497e+01 -37.425 < 2e-16 \*\*\*  
## waterfront 5.656e+05 2.059e+04 27.463 < 2e-16 \*\*\*  
## bedrooms -3.394e+04 2.191e+03 -15.495 < 2e-16 \*\*\*  
## bathrooms 4.063e+04 3.642e+03 11.154 < 2e-16 \*\*\*  
## zipcode -6.284e+02 3.838e+01 -16.375 < 2e-16 \*\*\*  
## long -1.850e+05 1.491e+04 -12.414 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 203900 on 15766 degrees of freedom  
## Multiple R-squared: 0.6878, Adjusted R-squared: 0.6876   
## F-statistic: 3473 on 10 and 15766 DF, p-value: < 2.2e-16

predictionXGen <- predict(modelXGen,interval='prediction',newdata = test)  
test\_rmseXGen = sqrt(sum((predictionXGen - test$price)^2)/nrow(test))  
errors <- predictionXGen[,'fit'] - test$price  
hist(errors,col=(c("gold","darkgreen")))



rel\_change = 1 - ((test$price - abs(errors)) / test$price)  
##Now the percentage of cases with less than 25% error.  
pred25 = table(rel\_change<0.25)["TRUE"] / nrow(test)  
pred25

## TRUE   
## 0.6201165

cat("\nConclusion:percent of data having less than 25% error:",pred25)

##   
## Conclusion:percent of data having less than 25% error: 0.6201165