checkNaFunction <- function(HousePrice){  
naColumns <- c()  
#checking NA for each columns  
for(i in 1:ncol(HousePrice)) {  
 #cat(sprintf("Checking NA: %s \n", colnames(HousePrice)[i]))  
 if(length(which(is.na(HousePrice[,i]))) > 0){  
 #cat(sprintf("There is NA: %s \n" , colnames(HousePrice)[i]))  
 naColumns <- c(naColumns, colnames(HousePrice)[i])  
 }  
}  
return(naColumns)  
}

analysis <- function(HousePrice, i, labels){  
 plot(HousePrice$price~HousePrice[,i],main = labels[1],xlab = labels[2],ylab = labels[3],col.main='Blue',col.axis='Gray',col.lab = 'Gray',col = 'Gray')  
   
 hist(HousePrice[,i],main = labels[1],xlab = labels[2],ylab = labels[3],col.main='Blue',col.axis='Gray',col.lab = 'Gray',col = 'Gray')  
  
 boxplot(HousePrice[,i],main = labels[1],col.main='Blue',col.axis='Gray',col.lab = 'Gray',col = 'Gray')  
   
 cor(HousePrice[,i],HousePrice$price)  
}

# Data Importing And Cleaning

HousePrice <- read.csv("file:///G:/Ryerson-BigData/capstone-R/CKME-136/data/kc\_house\_data.csv")  
head(HousePrice)

## id date price bedrooms bathrooms sqft\_living  
## 1 7129300520 20141013T000000 221900 3 1.00 1180  
## 2 6414100192 20141209T000000 538000 3 2.25 2570  
## 3 5631500400 20150225T000000 180000 2 1.00 770  
## 4 2487200875 20141209T000000 604000 4 3.00 1960  
## 5 1954400510 20150218T000000 510000 3 2.00 1680  
## 6 7237550310 20140512T000000 1225000 4 4.50 5420  
## sqft\_lot floors waterfront view condition grade sqft\_above sqft\_basement  
## 1 5650 1 0 0 3 7 1180 0  
## 2 7242 2 0 0 3 7 2170 400  
## 3 10000 1 0 0 3 6 770 0  
## 4 5000 1 0 0 5 7 1050 910  
## 5 8080 1 0 0 3 8 1680 0  
## 6 101930 1 0 0 3 11 3890 1530  
## yr\_built yr\_renovated zipcode lat long sqft\_living15 sqft\_lot15  
## 1 1955 0 98178 47.5112 -122.257 1340 5650  
## 2 1951 1991 98125 47.7210 -122.319 1690 7639  
## 3 1933 0 98028 47.7379 -122.233 2720 8062  
## 4 1965 0 98136 47.5208 -122.393 1360 5000  
## 5 1987 0 98074 47.6168 -122.045 1800 7503  
## 6 2001 0 98053 47.6561 -122.005 4760 101930

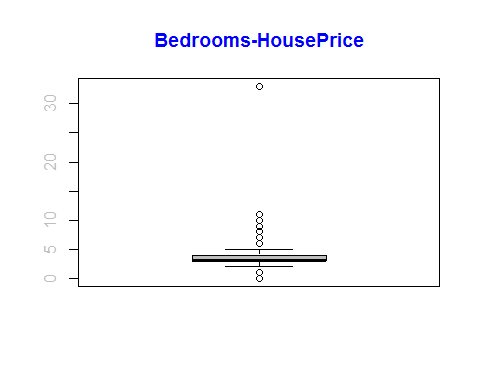
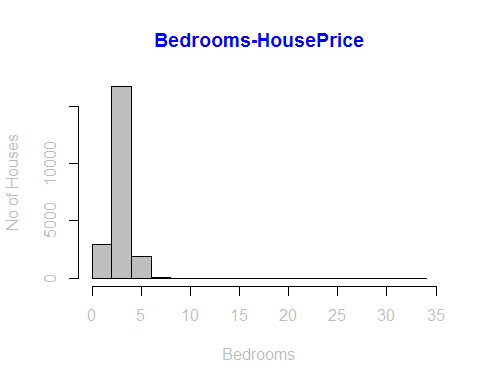
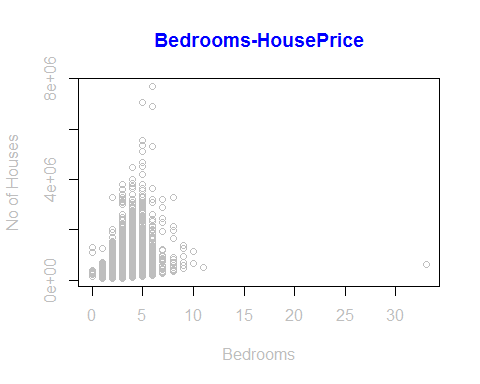
colnames(HousePrice)

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

naColumns <- checkNaFunction(HousePrice)  
if(length(naColumns)>0){  
 cat("Found NA Colums:")  
 for(i in 1:length(naColumns)) {  
 cat(sprintf("%s,", colnames(HousePrice)[i]))  
 }  
}  
# Here we conclude that this data does not hold any column with NA.

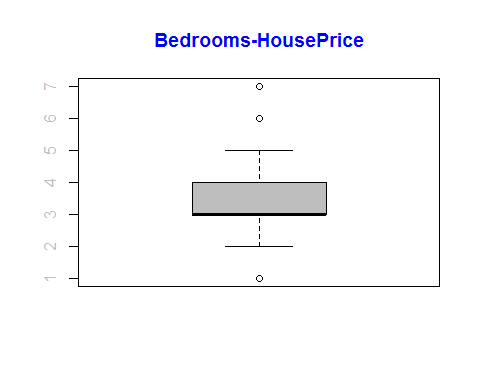
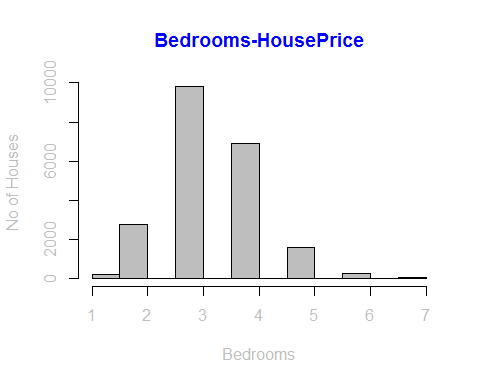
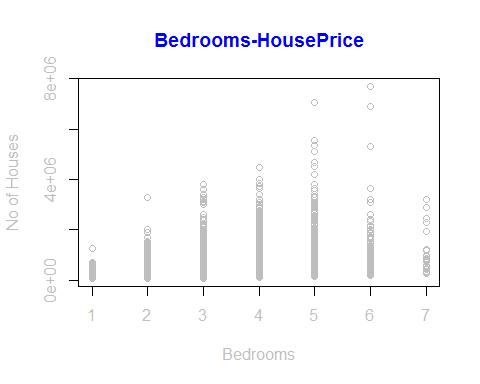
# BEDROOM ANALYSIS

analysis(HousePrice,4,c('Bedrooms-HousePrice','Bedrooms', 'No of Houses'))



## [1] 0.3083496

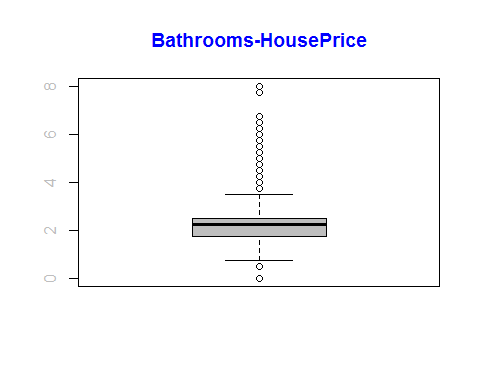
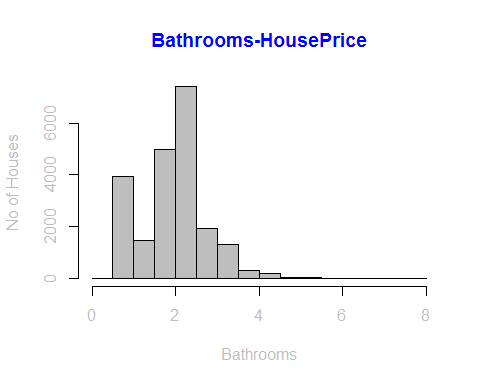
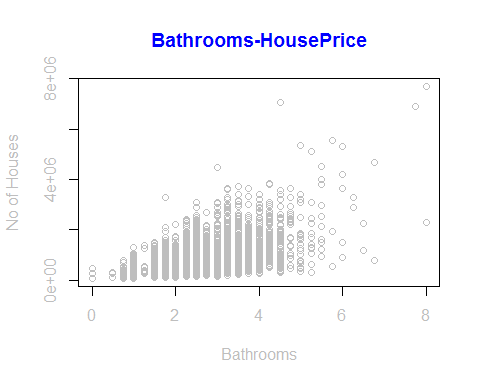
#\*\*\*\*\*\*\*Removing the outliers  
#Since more than 7 bedrooms are very rare.Also it's the outlier for my model.  
#I have removed the outlier data.  
HousePrice<-subset(HousePrice,bedrooms>=1 & bedrooms<=7)  
#\*\*\*\*\*\*\*Once we removed the outliers, again get the analysis  
analysis(HousePrice,4,c('Bedrooms-HousePrice','Bedrooms', 'No of Houses'))



## [1] 0.3156734

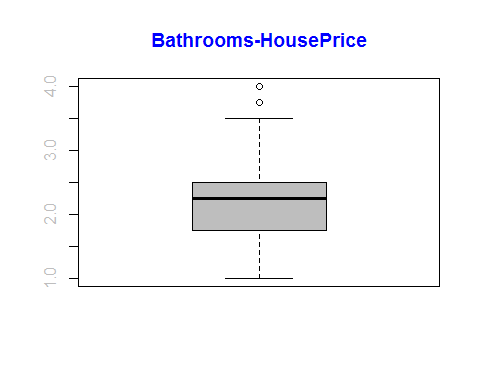
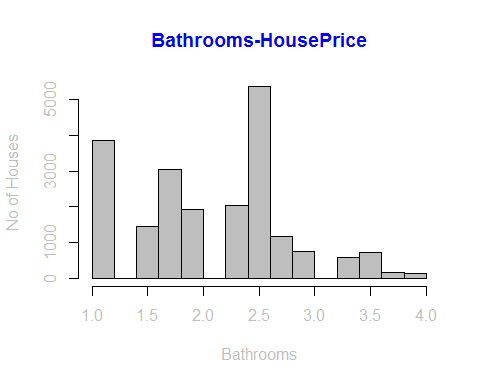
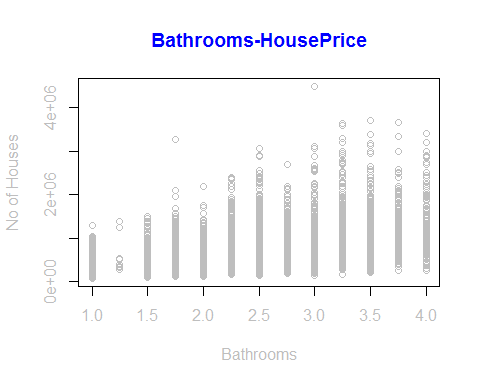
# BATHROOM ANALYSIS

analysis(HousePrice,5,c('Bathrooms-HousePrice','Bathrooms', 'No of Houses'))



## [1] 0.5259342

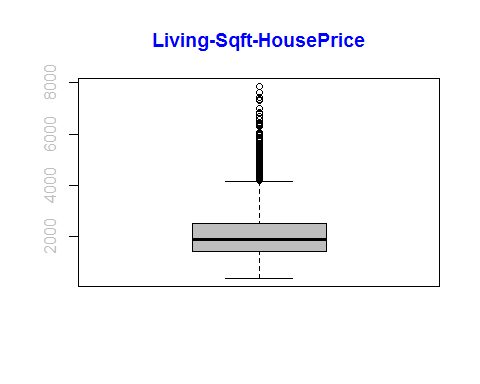
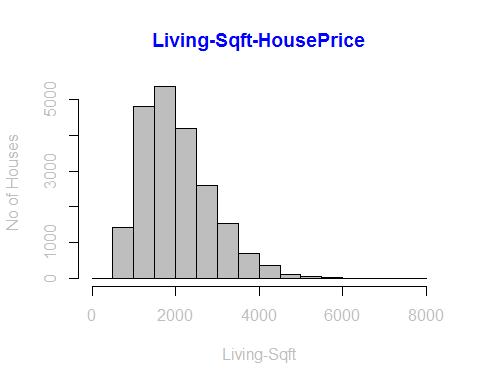
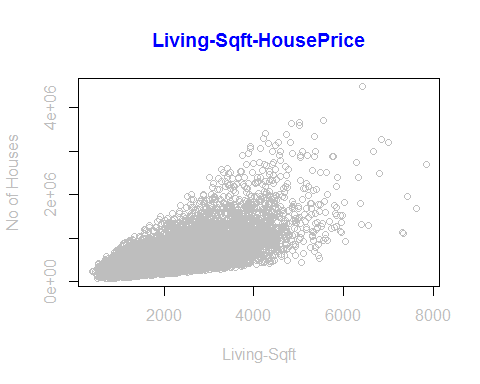
#\*\*\*\*\*\*\*Removing the outliers  
#More than 4 bathrooms are very rare in this data.So I am removing it.  
HousePrice<-subset(HousePrice,bathrooms>=1 & bathrooms<=4)  
analysis(HousePrice,5,c('Bathrooms-HousePrice','Bathrooms', 'No of Houses'))



## [1] 0.475159

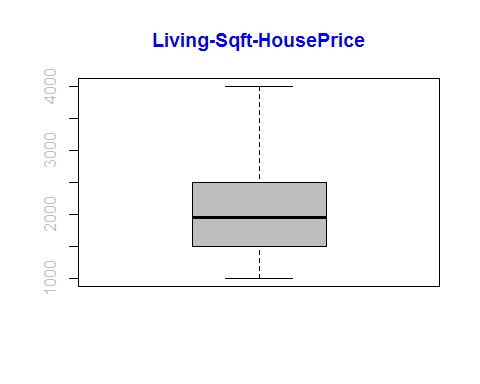
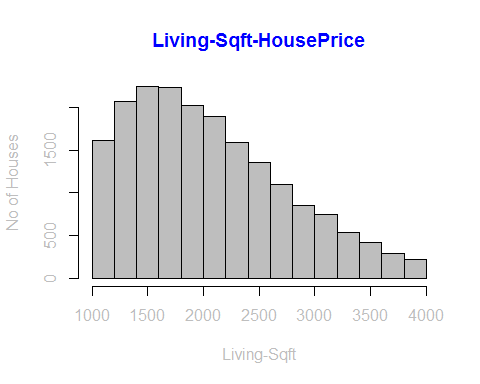
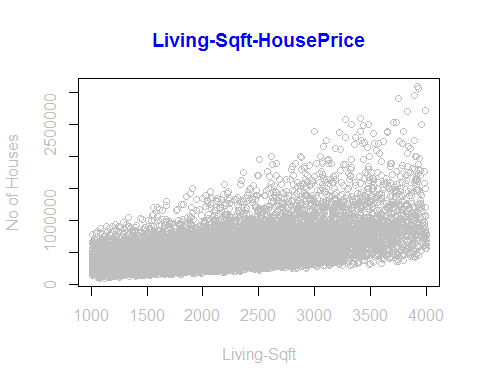
# SQFT LIVING ANALYSIS

analysis(HousePrice,6,c('Living-Sqft-HousePrice','Living-Sqft', 'No of Houses'))



## [1] 0.6701029

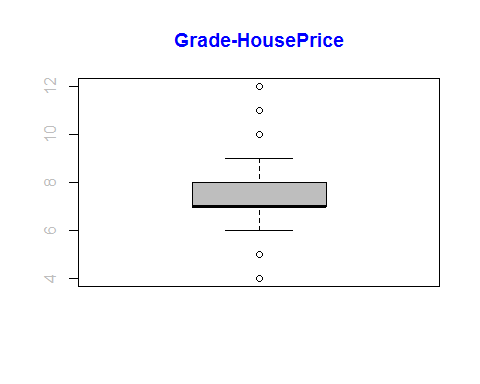
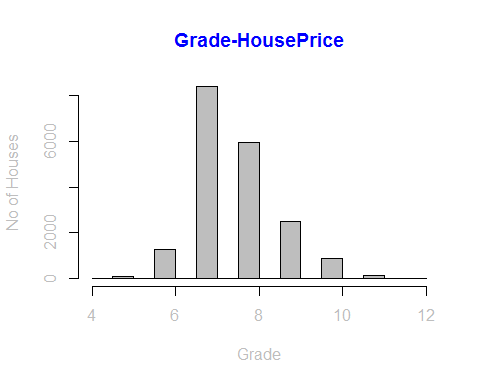
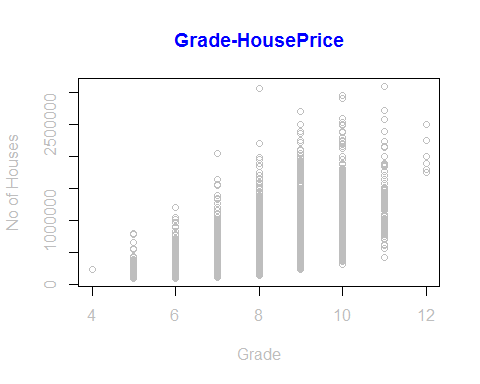
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_living >1000 & sqft\_living<=4000)  
analysis(HousePrice,6,c('Living-Sqft-HousePrice','Living-Sqft', 'No of Houses'))



## [1] 0.5938015

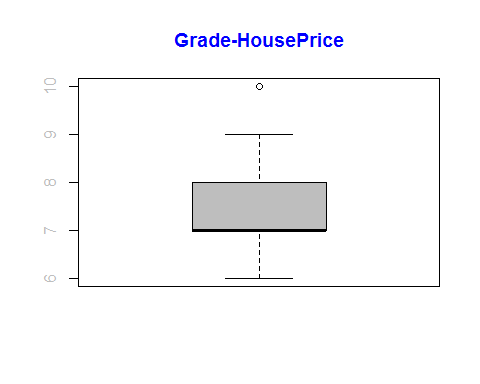
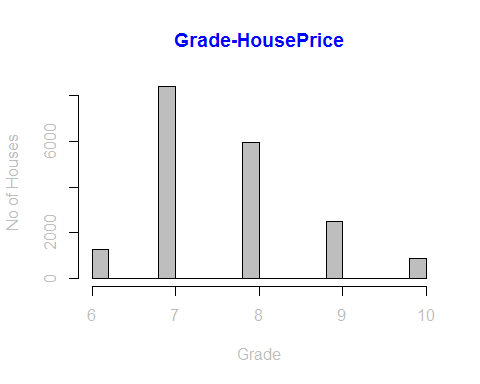
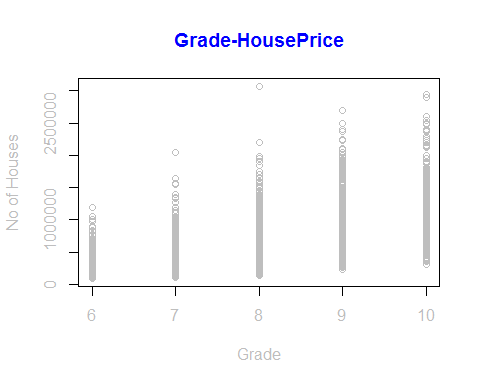
## GRADE ANALYSIS

analysis(HousePrice,12,c('Grade-HousePrice','Grade', 'No of Houses'))



## [1] 0.6106929

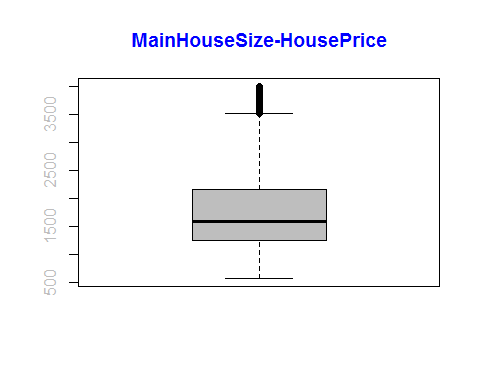
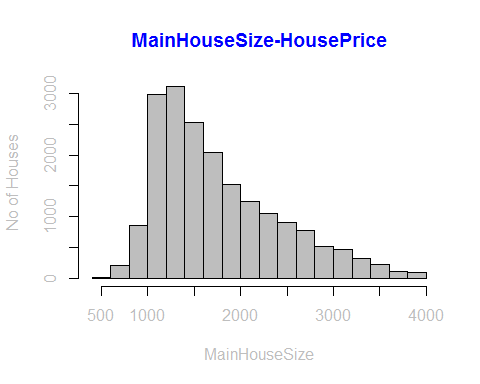
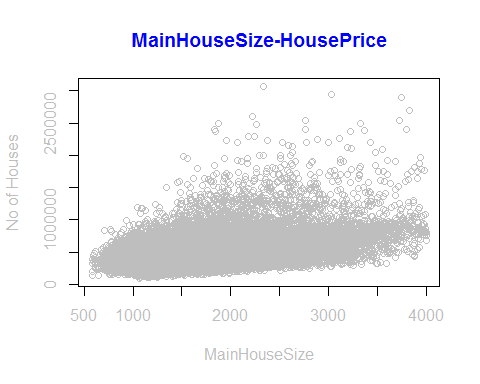
#\*\*\*\*\*\*\*Removing the outliers  
#Most of the houses grades are between 6-10   
HousePrice<-subset(HousePrice,grade >= 6 & grade<=10)  
analysis(HousePrice,12,c('Grade-HousePrice','Grade', 'No of Houses'))



## [1] 0.592697

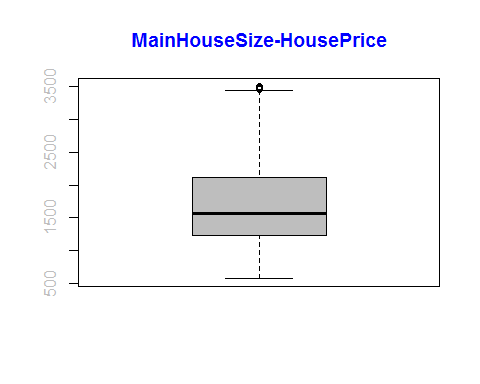
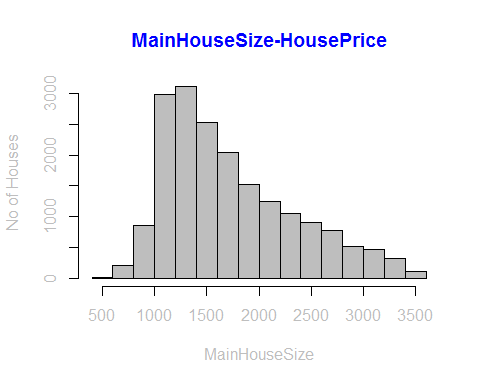
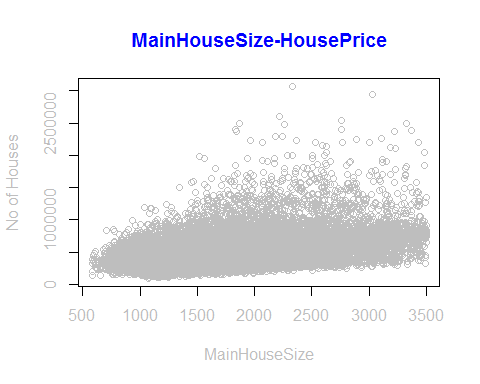
# SQFT\_ABOVE ANALYSIS

analysis(HousePrice,13,c('MainHouseSize-HousePrice','MainHouseSize', 'No of Houses'))



## [1] 0.4553313

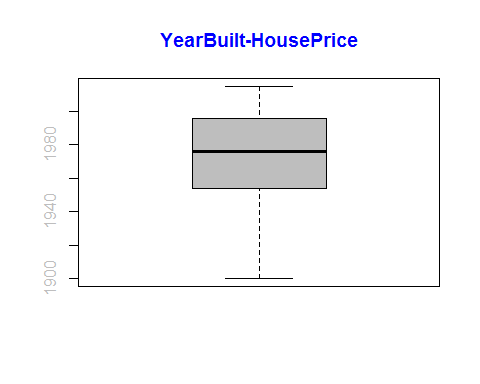
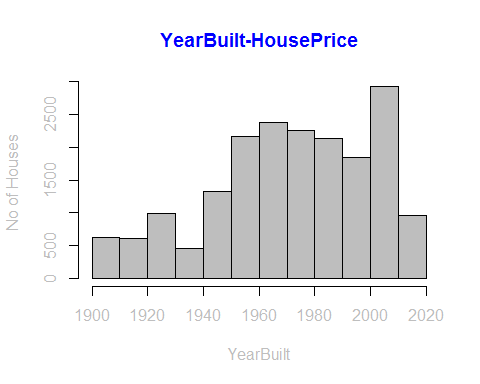
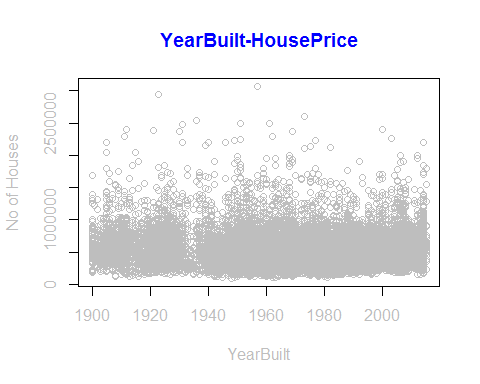
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_above >=500 & sqft\_above<=3500)  
analysis(HousePrice,13,c('MainHouseSize-HousePrice','MainHouseSize', 'No of Houses'))



## [1] 0.4254812

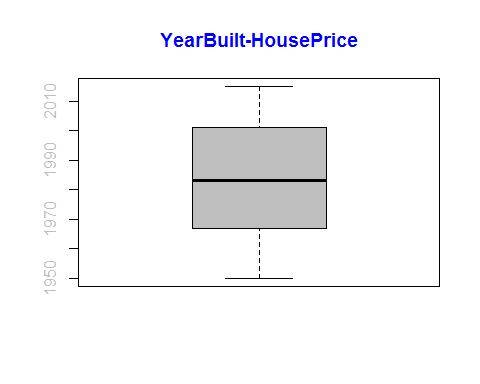
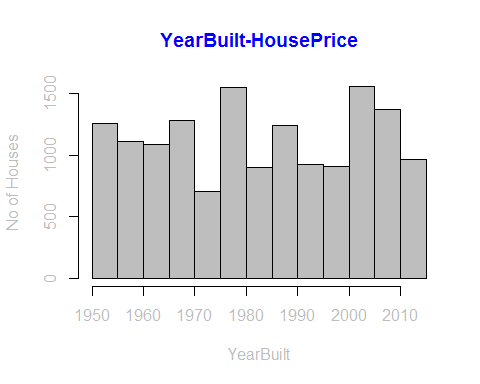
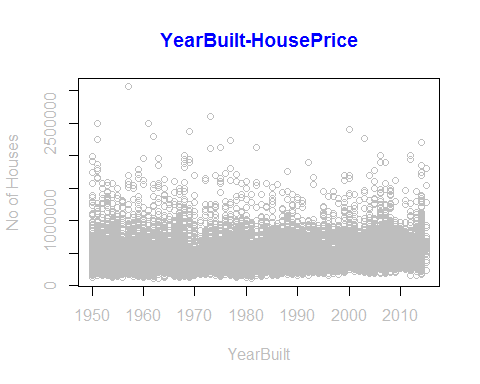
# YR\_BUILT ANALYSIS

analysis(HousePrice,15,c('YearBuilt-HousePrice','YearBuilt', 'No of Houses'))



## [1] -0.07805272

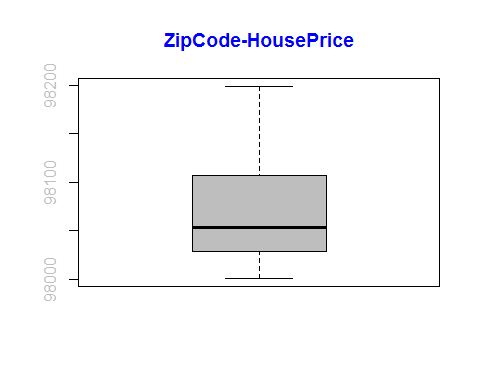
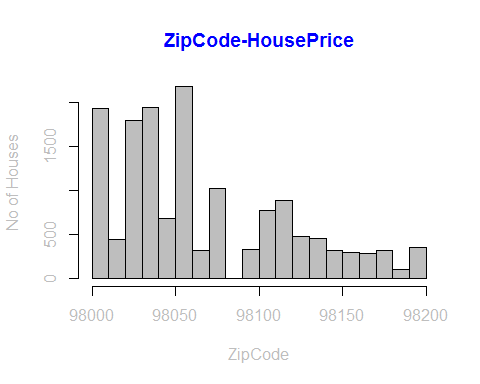
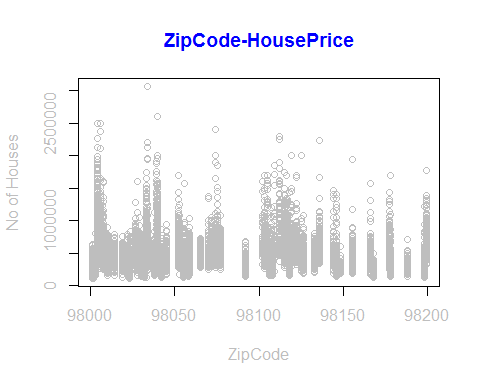
#\*\*\*\*\*\*\*Removing the outliers  
#In our data some records are too old..I just removed that data from my model.  
#Because It doesn't make any sense to keep more than 100 years house in our model  
HousePrice<-subset(HousePrice,yr\_built>=1950& yr\_built<=2015)  
analysis(HousePrice,15,c('YearBuilt-HousePrice','YearBuilt', 'No of Houses'))



## [1] 0.08555947

# ZIPCODE ANALYSIS

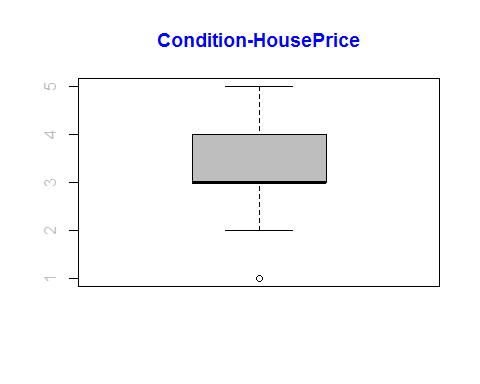
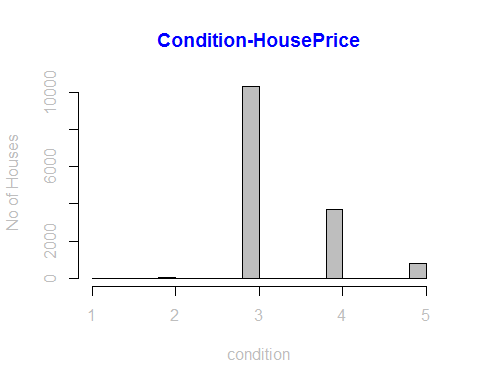
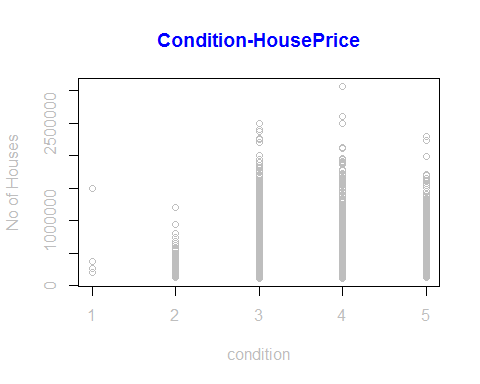
analysis(HousePrice,17,c('ZipCode-HousePrice','ZipCode', 'No of Houses'))



## [1] -0.02647096

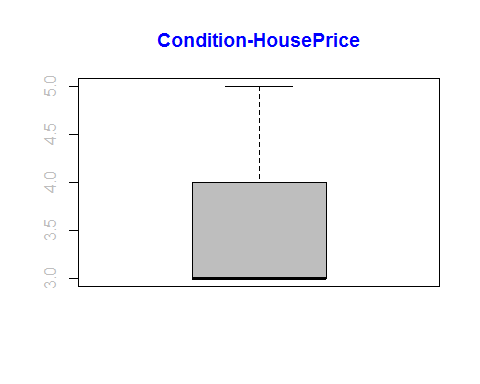
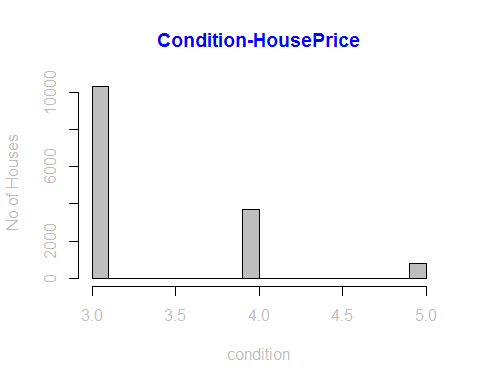
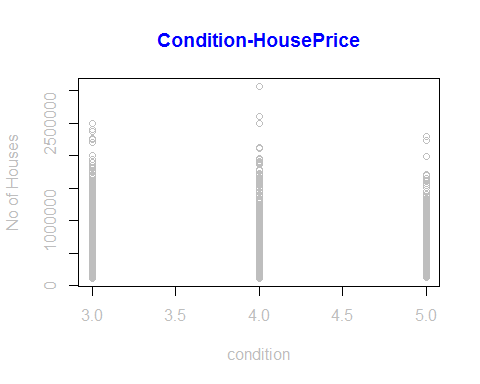
## CONDITION aNALYSIS

analysis(HousePrice,11,c('Condition-HousePrice','condition', 'No of Houses'))



## [1] 0.0300221

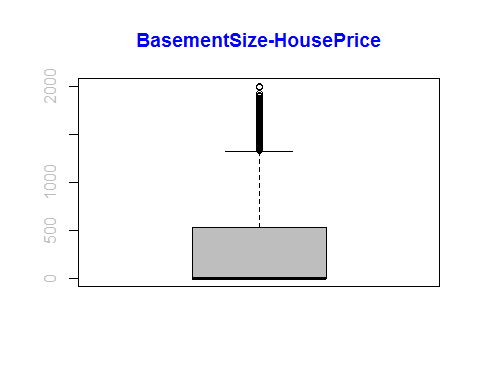
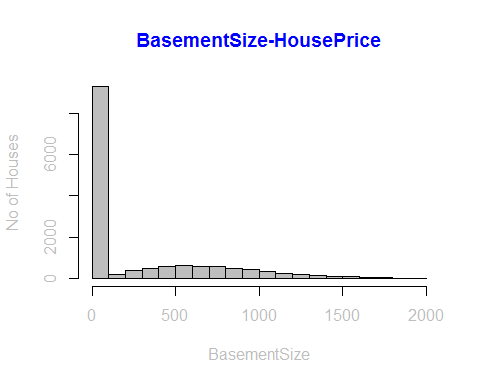
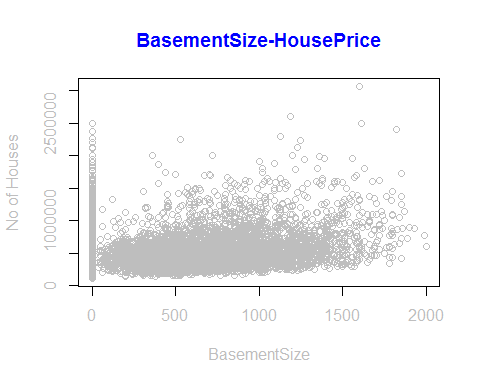
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,condition>=3& condition<=5)  
analysis(HousePrice,11,c('Condition-HousePrice','condition', 'No of Houses'))



## [1] 0.02442002

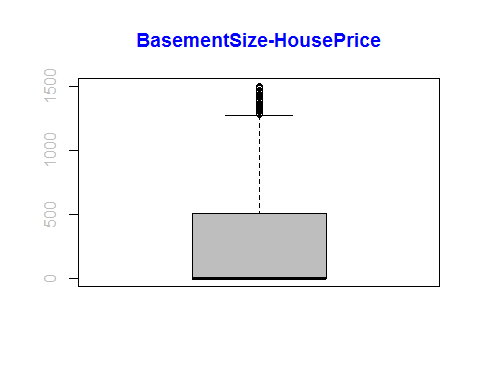
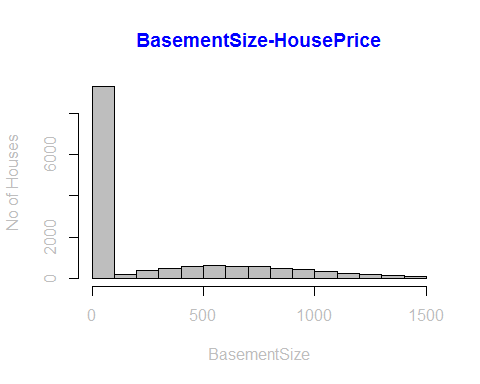
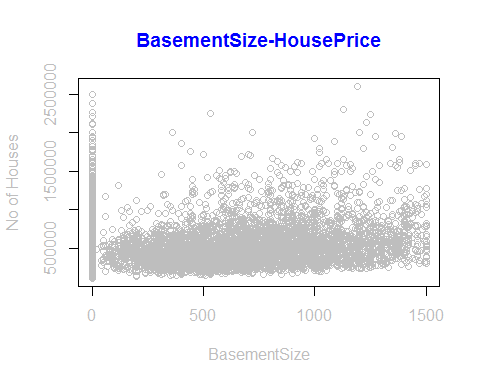
## SQFT\_BASEMENT ANALYSIS

analysis(HousePrice,14,c('BasementSize-HousePrice','BasementSize', 'No of Houses'))



## [1] 0.2254519

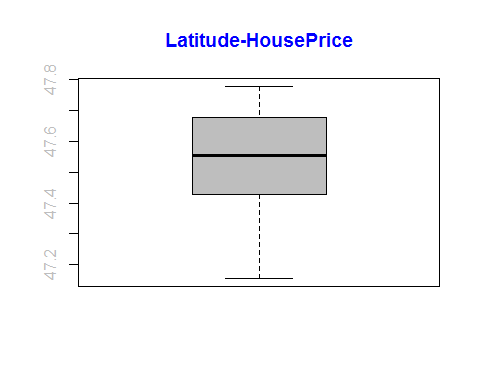
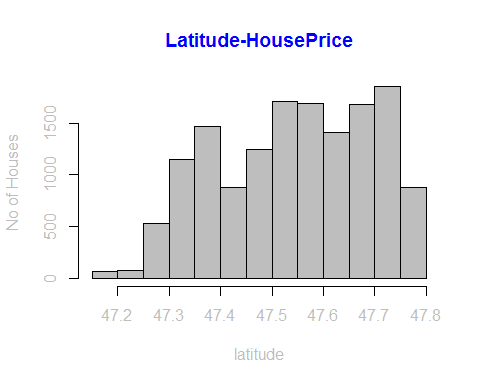
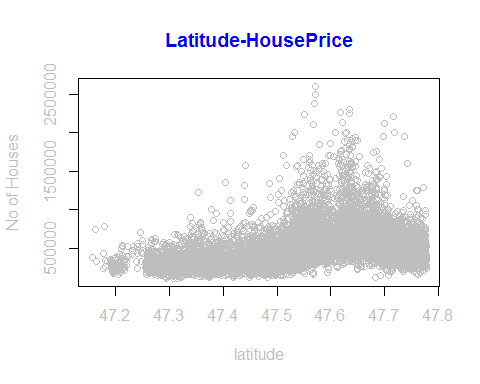
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_basement >=0 & sqft\_basement<=1500)  
analysis(HousePrice,14,c('BasementSize-HousePrice','BasementSize', 'No of Houses'))



## [1] 0.1879221

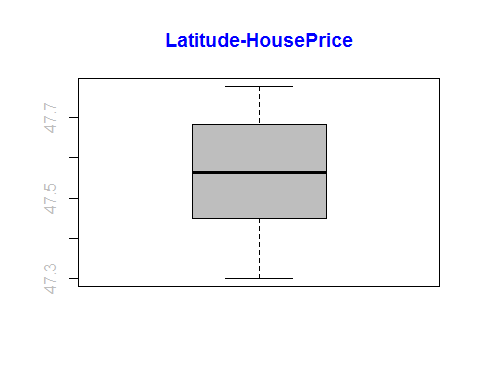
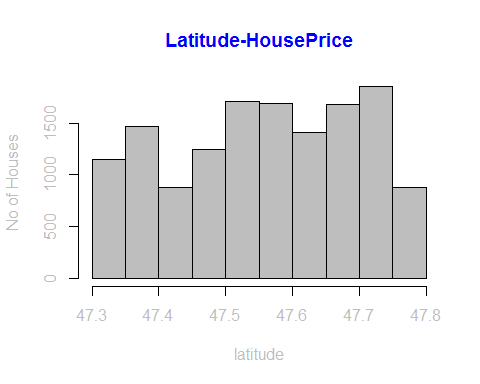
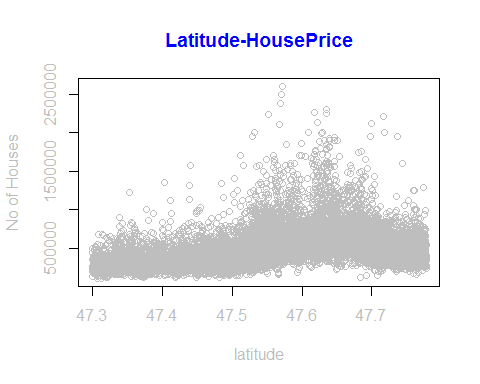
## LAT ANALYSIS

analysis(HousePrice,18,c('Latitude-HousePrice','latitude', 'No of Houses'))



## [1] 0.4161934

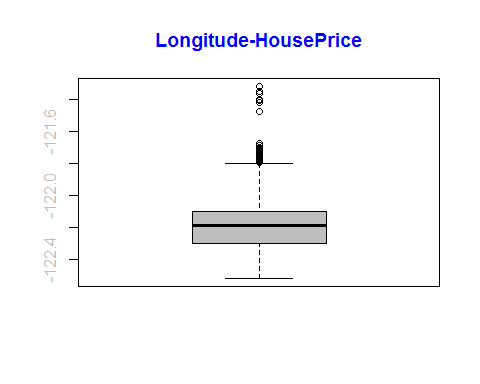
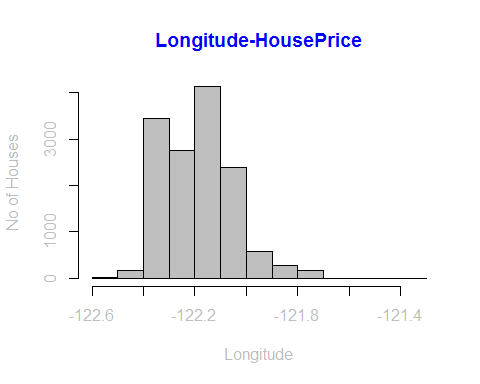
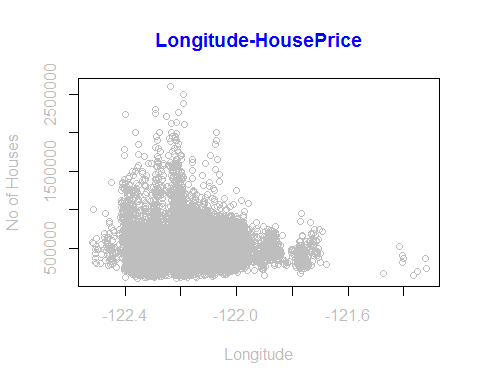
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,lat>=47.3)  
analysis(HousePrice,18,c('Latitude-HousePrice','latitude', 'No of Houses'))



## [1] 0.3848454

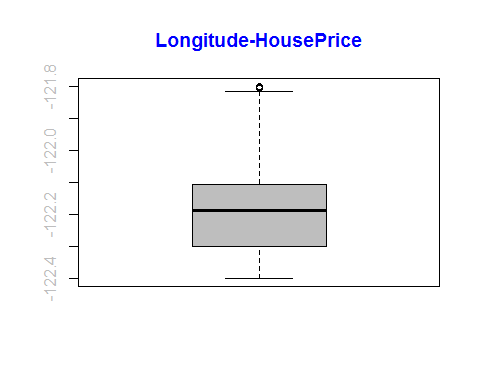
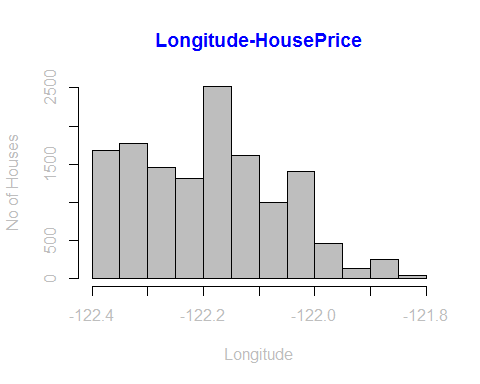
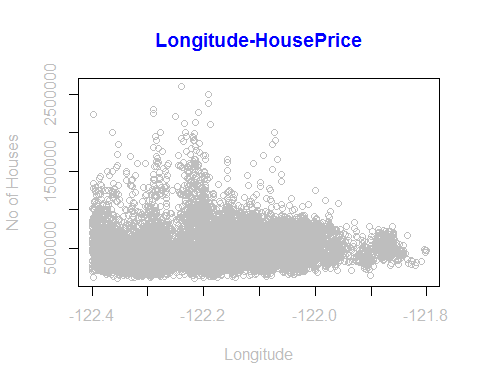
## LONG ANALYSIS

analysis(HousePrice,19,c('Longitude-HousePrice','Longitude', 'No of Houses'))



## [1] 0.03740839

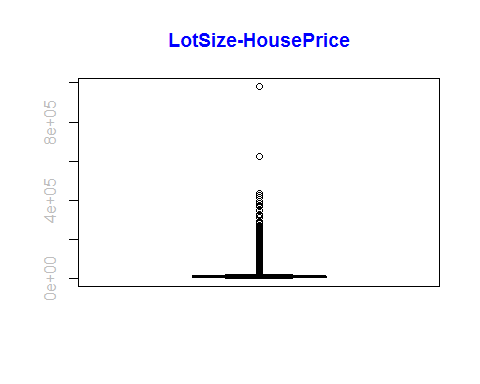
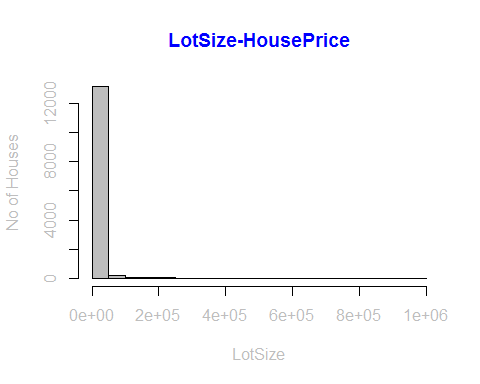
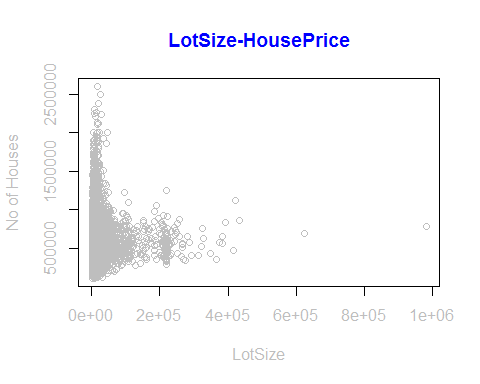
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,long>=-122.4 & long < -121.8)  
analysis(HousePrice,19,c('Longitude-HousePrice','Longitude', 'No of Houses'))



## [1] 0.0746599

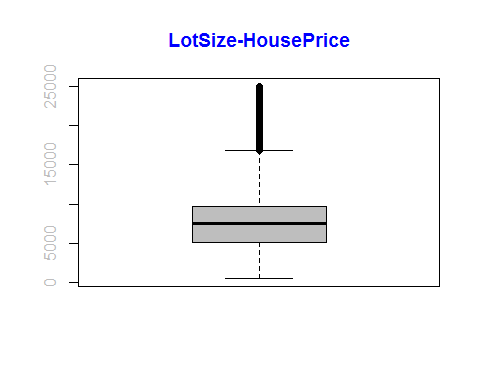
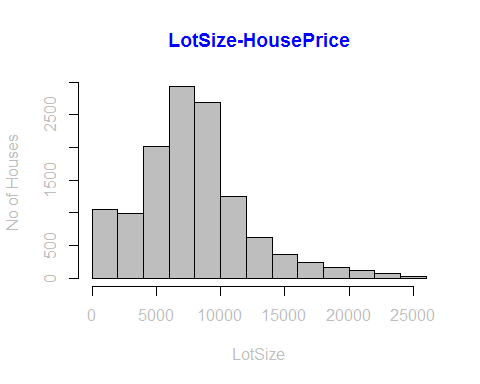
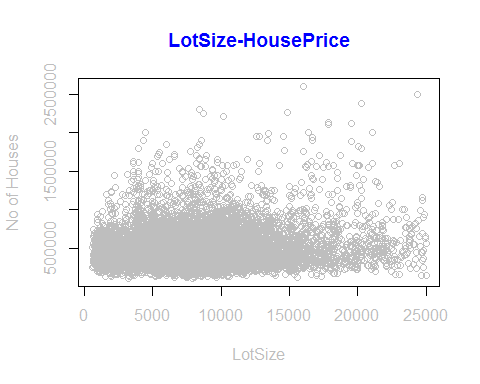
## SQFT\_LOT ANALYSIS

analysis(HousePrice,7,c('LotSize-HousePrice','LotSize', 'No of Houses'))



## [1] 0.09390182

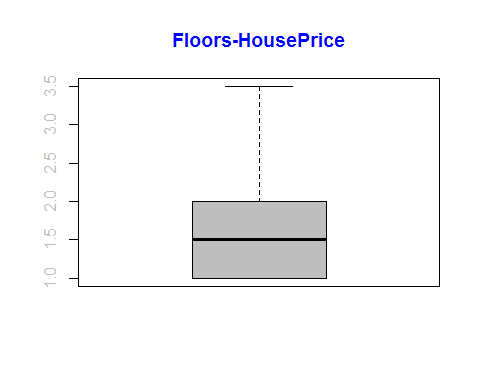
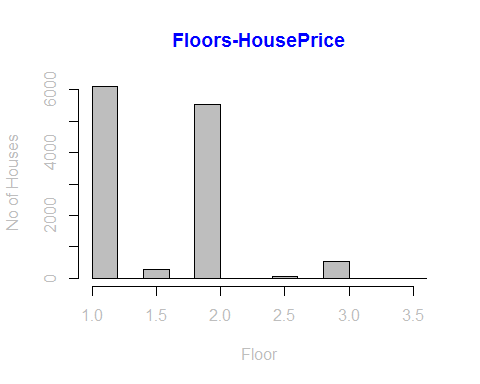
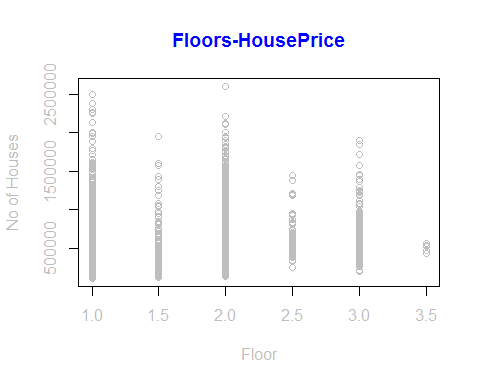
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_lot>=0 & sqft\_lot<=25000)  
analysis(HousePrice,7,c('LotSize-HousePrice','LotSize', 'No of Houses'))



## [1] 0.1228599

## FLOOR ANALYSIS

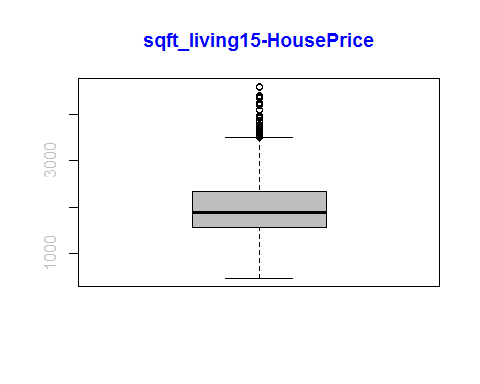
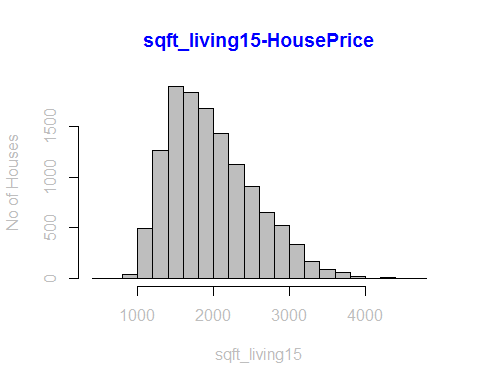
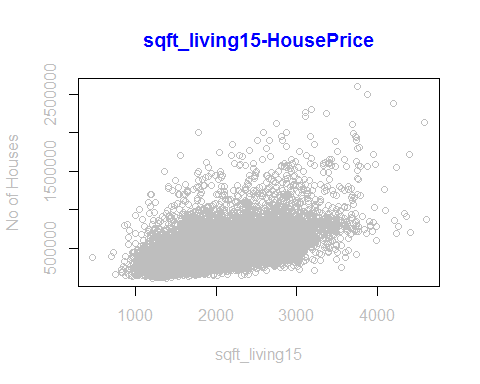
analysis(HousePrice,8,c('Floors-HousePrice','Floor', 'No of Houses'))



## [1] 0.189869

## SQFT\_LIVING15 ANALYSIS

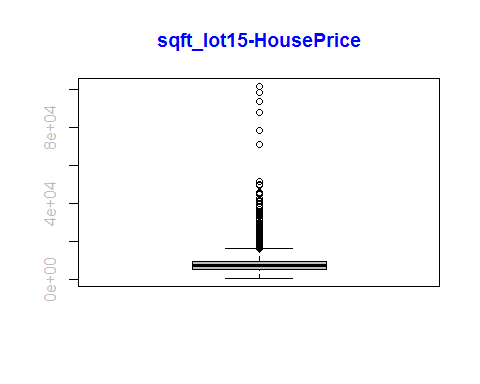
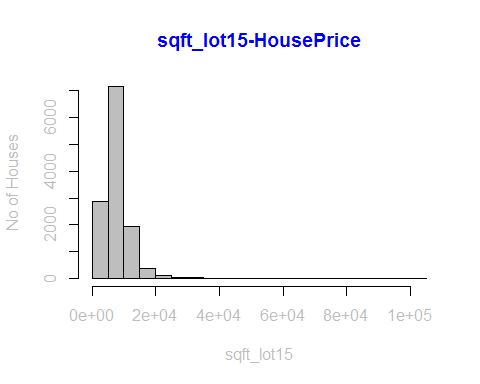
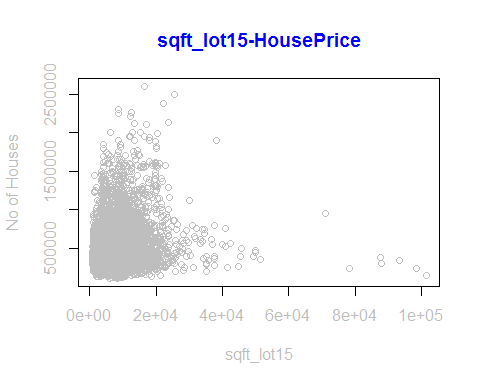
analysis(HousePrice,20,c('sqft\_living15-HousePrice','sqft\_living15', 'No of Houses'))



## [1] 0.5377631

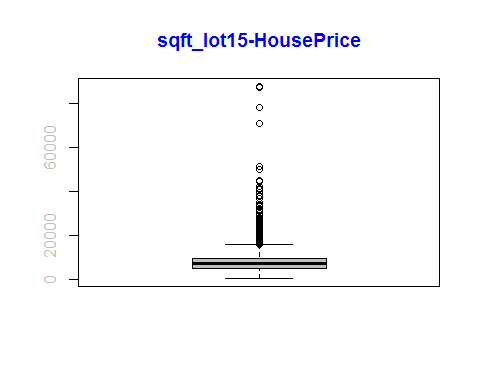
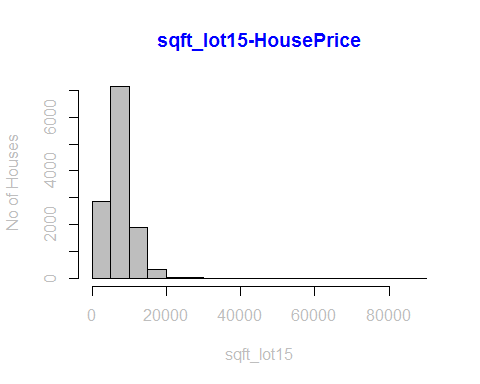
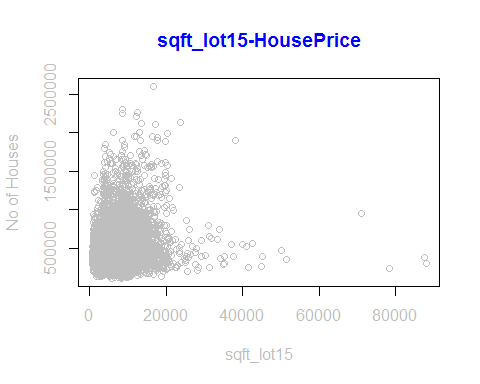
## SQFT\_LOT15 ANALYSIS

analysis(HousePrice,21,c('sqft\_lot15-HousePrice','sqft\_lot15', 'No of Houses'))



## [1] 0.1155467

#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_lot15>=0 & sqft\_lot<=20000)  
analysis(HousePrice,21,c('sqft\_lot15-HousePrice','sqft\_lot15', 'No of Houses'))



## [1] 0.1046768

## Correlation among all the variables

#\*\*\*\*\*\*\*Only sqft\_living & sqft\_above,sqft\_living & grade,sqft\_living & bathrooms have good correlation between them  
HousePrice$date <- NULL  
cor(HousePrice)

## id price bedrooms bathrooms sqft\_living  
## id 1.00000000 0.03088511 -0.00248828 0.05248387 0.05591911  
## price 0.03088511 1.00000000 0.21316110 0.40694539 0.58753189  
## bedrooms -0.00248828 0.21316110 1.00000000 0.35967366 0.54595425  
## bathrooms 0.05248387 0.40694539 0.35967366 1.00000000 0.61084492  
## sqft\_living 0.05591911 0.58753189 0.54595425 0.61084492 1.00000000  
## sqft\_lot -0.04881811 0.09879146 0.21379756 -0.15003858 0.21349737  
## floors 0.03142587 0.19599487 -0.01231716 0.48824267 0.21834143  
## waterfront -0.00601867 0.18723046 -0.02388155 0.02402709 0.04512995  
## view 0.04845104 0.32126248 0.04370921 0.10743868 0.20290469  
## condition -0.05442527 0.01080196 0.05075847 -0.20275433 -0.08704616  
## grade 0.06542157 0.62532384 0.17613009 0.51361808 0.62530895  
## sqft\_above 0.06310017 0.47583099 0.37950422 0.52925673 0.80769249  
## sqft\_basement -0.01105019 0.18482790 0.27249273 0.13661508 0.31810283  
## yr\_built 0.07163873 0.11886923 -0.02014115 0.56553240 0.25215473  
## yr\_renovated -0.01141468 0.10260750 0.01761968 0.01698352 0.02780418  
## zipcode -0.03344919 -0.06661544 -0.10383298 -0.08873701 -0.13655332  
## lat -0.01028660 0.39782088 -0.04362736 0.03452169 0.02941721  
## long 0.09490813 0.06021992 0.09157411 0.15541240 0.21179591  
## sqft\_living15 0.07487106 0.53192136 0.34578440 0.44219575 0.73234542  
## sqft\_lot15 -0.05544805 0.10467684 0.16855680 -0.13671006 0.17733937  
## sqft\_lot floors waterfront view condition  
## id -0.04881811 0.03142587 -0.006018670 0.04845104 -0.05442527  
## price 0.09879146 0.19599487 0.187230461 0.32126248 0.01080196  
## bedrooms 0.21379756 -0.01231716 -0.023881549 0.04370921 0.05075847  
## bathrooms -0.15003858 0.48824267 0.024027088 0.10743868 -0.20275433  
## sqft\_living 0.21349737 0.21834143 0.045129953 0.20290469 -0.08704616  
## sqft\_lot 1.00000000 -0.46764775 0.061281460 0.10171693 0.27240654  
## floors -0.46764775 1.00000000 0.020419644 -0.01755120 -0.36800514  
## waterfront 0.06128146 0.02041964 1.000000000 0.34042155 0.01009095  
## view 0.10171693 -0.01755120 0.340421549 1.00000000 0.03451890  
## condition 0.27240654 -0.36800514 0.010090951 0.03451890 1.00000000  
## grade 0.01799467 0.40495046 0.058898054 0.18464079 -0.19021868  
## sqft\_above 0.10812840 0.46766878 0.019014619 0.06247316 -0.19571109  
## sqft\_basement 0.17132458 -0.39896785 0.042393068 0.22760950 0.17395516  
## yr\_built -0.47268086 0.72279773 -0.025271092 -0.08730803 -0.45609638  
## yr\_renovated 0.04836387 -0.03403060 0.050423835 0.06473431 -0.04580786  
## zipcode -0.20648742 0.02202050 0.049153202 0.11697692 -0.09996766  
## lat -0.05925894 0.05026221 -0.007680641 0.01912452 -0.04700562  
## long 0.15246300 0.07677540 -0.010330764 -0.10532559 -0.04109049  
## sqft\_living15 0.26717897 0.16001569 0.063053624 0.22038305 -0.07723497  
## sqft\_lot15 0.80071360 -0.40438723 0.101403098 0.10986952 0.24966404  
## grade sqft\_above sqft\_basement yr\_built  
## id 0.0654215660 0.063100165 -0.0110501872 0.07163873  
## price 0.6253238358 0.475830987 0.1848279021 0.11886923  
## bedrooms 0.1761300924 0.379504224 0.2724927258 -0.02014115  
## bathrooms 0.5136180759 0.529256730 0.1366150794 0.56553240  
## sqft\_living 0.6253089478 0.807692487 0.3181028274 0.25215473  
## sqft\_lot 0.0179946663 0.108128401 0.1713245799 -0.47268086  
## floors 0.4049504567 0.467668775 -0.3989678500 0.72279773  
## waterfront 0.0588980541 0.019014619 0.0423930680 -0.02527109  
## view 0.1846407902 0.062473162 0.2276095048 -0.08730803  
## condition -0.1902186768 -0.195711088 0.1739551648 -0.45609638  
## grade 1.0000000000 0.629292059 -0.0008512015 0.41391909  
## sqft\_above 0.6292920588 1.000000000 -0.3020484503 0.43178138  
## sqft\_basement -0.0008512015 -0.302048450 1.0000000000 -0.28659200  
## yr\_built 0.4139190905 0.431781375 -0.2865920034 1.00000000  
## yr\_renovated 0.0137545251 0.005527641 0.0360665829 -0.14488568  
## zipcode -0.0963141760 -0.207857824 0.1134416369 -0.10071641  
## lat 0.1244168337 -0.030479955 0.0965731536 -0.05718985  
## long 0.1212933902 0.346947946 -0.2154373008 0.25864663  
## sqft\_living15 0.6095840261 0.677956453 0.0939590371 0.21331396  
## sqft\_lot15 0.0242929122 0.088065155 0.1451238319 -0.41247570  
## yr\_renovated zipcode lat long  
## id -0.011414678 -0.03344919 -0.010286601 0.09490813  
## price 0.102607501 -0.06661544 0.397820880 0.06021992  
## bedrooms 0.017619683 -0.10383298 -0.043627362 0.09157411  
## bathrooms 0.016983524 -0.08873701 0.034521688 0.15541240  
## sqft\_living 0.027804177 -0.13655332 0.029417207 0.21179591  
## sqft\_lot 0.048363871 -0.20648742 -0.059258942 0.15246300  
## floors -0.034030600 0.02202050 0.050262205 0.07677540  
## waterfront 0.050423835 0.04915320 -0.007680641 -0.01033076  
## view 0.064734312 0.11697692 0.019124515 -0.10532559  
## condition -0.045807864 -0.09996766 -0.047005621 -0.04109049  
## grade 0.013754525 -0.09631418 0.124416834 0.12129339  
## sqft\_above 0.005527641 -0.20785782 -0.030479955 0.34694795  
## sqft\_basement 0.036066583 0.11344164 0.096573154 -0.21543730  
## yr\_built -0.144885684 -0.10071641 -0.057189850 0.25864663  
## yr\_renovated 1.000000000 0.04008558 0.021904703 -0.03691168  
## zipcode 0.040085583 1.00000000 0.185270334 -0.54354459  
## lat 0.021904703 0.18527033 1.000000000 -0.08224426  
## long -0.036911680 -0.54354459 -0.082244257 1.00000000  
## sqft\_living15 0.002850830 -0.25878003 0.012087716 0.33806073  
## sqft\_lot15 0.060629185 -0.16726839 -0.043272796 0.13205797  
## sqft\_living15 sqft\_lot15  
## id 0.07487106 -0.05544805  
## price 0.53192136 0.10467684  
## bedrooms 0.34578440 0.16855680  
## bathrooms 0.44219575 -0.13671006  
## sqft\_living 0.73234542 0.17733937  
## sqft\_lot 0.26717897 0.80071360  
## floors 0.16001569 -0.40438723  
## waterfront 0.06305362 0.10140310  
## view 0.22038305 0.10986952  
## condition -0.07723497 0.24966404  
## grade 0.60958403 0.02429291  
## sqft\_above 0.67795645 0.08806515  
## sqft\_basement 0.09395904 0.14512383  
## yr\_built 0.21331396 -0.41247570  
## yr\_renovated 0.00285083 0.06062919  
## zipcode -0.25878003 -0.16726839  
## lat 0.01208772 -0.04327280  
## long 0.33806073 0.13205797  
## sqft\_living15 1.00000000 0.25196255  
## sqft\_lot15 0.25196255 1.00000000

# Linear Model using stepAIC method

lm <- lm(formula = price~.,data=HousePrice)  
install.packages("MASS", repos = "http://cran.us.r-project.org")

## Installing package into 'C:/Users/Geeta/Documents/R/win-library/3.3'  
## (as 'lib' is unspecified)

## package 'MASS' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Geeta\AppData\Local\Temp\RtmpYLtRg4\downloaded\_packages

library(MASS)

## Warning: package 'MASS' was built under R version 3.3.3

stepb<-stepAIC(lm,direction="backward",trace=TRUE)

## Start: AIC=291869.3  
## price ~ id + 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=291869.3  
## price ~ id + 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  
## - sqft\_lot15 1 5.9838e+07 2.4099e+14 291867  
## <none> 2.4099e+14 291869  
## - id 1 4.7429e+10 2.4104e+14 291870  
## - sqft\_above 1 6.2808e+10 2.4106e+14 291871  
## - floors 1 9.4826e+10 2.4109e+14 291872  
## - sqft\_lot 1 5.9161e+11 2.4159e+14 291897  
## - bathrooms 1 9.0197e+11 2.4190e+14 291913  
## - long 1 9.9620e+11 2.4199e+14 291918  
## - bedrooms 1 1.0407e+12 2.4204e+14 291920  
## - yr\_built 1 1.8331e+12 2.4283e+14 291961  
## - sqft\_living15 1 2.1249e+12 2.4312e+14 291975  
## - yr\_renovated 1 2.5732e+12 2.4357e+14 291998  
## - zipcode 1 3.8762e+12 2.4487e+14 292064  
## - condition 1 5.4017e+12 2.4640e+14 292140  
## - waterfront 1 6.4653e+12 2.4746e+14 292193  
## - view 1 8.6588e+12 2.4965e+14 292302  
## - sqft\_living 1 8.8652e+12 2.4986e+14 292312  
## - grade 1 3.2559e+13 2.7355e+14 293428  
## - lat 1 7.3430e+13 3.1442e+14 295143  
##   
## Step: AIC=291867.3  
## price ~ id + bedrooms + bathrooms + sqft\_living + sqft\_lot +   
## floors + waterfront + view + condition + grade + sqft\_above +   
## yr\_built + yr\_renovated + zipcode + lat + long + sqft\_living15  
##   
## Df Sum of Sq RSS AIC  
## <none> 2.4099e+14 291867  
## - id 1 4.7370e+10 2.4104e+14 291868  
## - sqft\_above 1 6.2767e+10 2.4106e+14 291869  
## - floors 1 9.5122e+10 2.4109e+14 291870  
## - bathrooms 1 9.0228e+11 2.4190e+14 291911  
## - long 1 9.9700e+11 2.4199e+14 291916  
## - bedrooms 1 1.0406e+12 2.4204e+14 291918  
## - sqft\_lot 1 1.0924e+12 2.4209e+14 291921  
## - yr\_built 1 1.8346e+12 2.4283e+14 291959  
## - sqft\_living15 1 2.1360e+12 2.4313e+14 291974  
## - yr\_renovated 1 2.5745e+12 2.4357e+14 291996  
## - zipcode 1 3.8770e+12 2.4487e+14 292062  
## - condition 1 5.4078e+12 2.4640e+14 292139  
## - waterfront 1 6.5023e+12 2.4750e+14 292193  
## - view 1 8.6588e+12 2.4965e+14 292300  
## - sqft\_living 1 8.8712e+12 2.4987e+14 292310  
## - grade 1 3.2565e+13 2.7356e+14 293426  
## - lat 1 7.3430e+13 3.1442e+14 295141

summary(stepb)

##   
## Call:  
## lm(formula = price ~ id + bedrooms + bathrooms + sqft\_living +   
## sqft\_lot + floors + waterfront + view + condition + grade +   
## sqft\_above + yr\_built + yr\_renovated + zipcode + lat + long +   
## sqft\_living15, data = HousePrice)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -821031 -79724 -9459 59576 1381714   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.480e+06 2.704e+06 0.917 0.3590   
## id -6.905e-07 4.441e-07 -1.555 0.1200   
## bedrooms -1.521e+04 2.087e+03 -7.287 3.37e-13 \*\*\*  
## bathrooms 2.301e+04 3.391e+03 6.785 1.21e-11 \*\*\*  
## sqft\_living 1.044e+02 4.907e+00 21.276 < 2e-16 \*\*\*  
## sqft\_lot -3.432e+00 4.597e-01 -7.466 8.83e-14 \*\*\*  
## floors 8.349e+03 3.790e+03 2.203 0.0276 \*   
## waterfront 4.429e+05 2.431e+04 18.215 < 2e-16 \*\*\*  
## view 5.008e+04 2.383e+03 21.020 < 2e-16 \*\*\*  
## condition 4.223e+04 2.542e+03 16.611 < 2e-16 \*\*\*  
## grade 9.021e+04 2.213e+03 40.763 < 2e-16 \*\*\*  
## sqft\_above 8.146e+00 4.552e+00 1.790 0.0735 .   
## yr\_built -1.200e+03 1.240e+02 -9.675 < 2e-16 \*\*\*  
## yr\_renovated 5.901e+01 5.148e+00 11.462 < 2e-16 \*\*\*  
## zipcode -4.195e+02 2.982e+01 -14.065 < 2e-16 \*\*\*  
## lat 5.994e+05 9.792e+03 61.211 < 2e-16 \*\*\*  
## long -9.722e+04 1.363e+04 -7.133 1.04e-12 \*\*\*  
## sqft\_living15 3.949e+01 3.783e+00 10.440 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 140000 on 12297 degrees of freedom  
## Multiple R-squared: 0.6447, Adjusted R-squared: 0.6442   
## F-statistic: 1312 on 17 and 12297 DF, p-value: < 2.2e-16

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

## 'data.frame': 12315 obs. of 20 variables:  
## $ id : num 7.13e+09 6.41e+09 2.49e+09 1.95e+09 1.32e+09 ...  
## $ price : num 221900 538000 604000 510000 257500 ...  
## $ bedrooms : int 3 3 4 3 3 3 3 3 3 4 ...  
## $ bathrooms : num 1 2.25 3 2 2.25 1.5 1 2.5 1.75 3 ...  
## $ sqft\_living : int 1180 2570 1960 1680 1715 1060 1780 1890 1370 2950 ...  
## $ sqft\_lot : int 5650 7242 5000 8080 6819 9711 7470 6560 9680 5000 ...  
## $ floors : num 1 2 1 1 2 1 1 2 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 3 ...  
## $ condition : int 3 3 5 3 3 3 3 3 4 3 ...  
## $ grade : int 7 7 7 8 7 7 7 7 7 9 ...  
## $ sqft\_above : int 1180 2170 1050 1680 1715 1060 1050 1890 1370 1980 ...  
## $ sqft\_basement: int 0 400 910 0 0 0 730 0 0 970 ...  
## $ yr\_built : int 1955 1951 1965 1987 1995 1963 1960 2003 1977 1979 ...  
## $ yr\_renovated : int 0 1991 0 0 0 0 0 0 0 0 ...  
## $ zipcode : int 98178 98125 98136 98074 98003 98198 98146 98038 98074 98126 ...  
## $ lat : num 47.5 47.7 47.5 47.6 47.3 ...  
## $ long : num -122 -122 -122 -122 -122 ...  
## $ sqft\_living15: int 1340 1690 1360 1800 2238 1650 1780 2390 1370 2140 ...  
## $ sqft\_lot15 : int 5650 7639 5000 7503 6819 9711 8113 7570 10208 4000 ...

sapply(HousePrice, is.numeric)

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

HousePrices <- HousePrice[ , sapply(HousePrice, is.numeric)]  
cor(HousePrices)

## id price bedrooms bathrooms sqft\_living  
## id 1.00000000 0.03088511 -0.00248828 0.05248387 0.05591911  
## price 0.03088511 1.00000000 0.21316110 0.40694539 0.58753189  
## bedrooms -0.00248828 0.21316110 1.00000000 0.35967366 0.54595425  
## bathrooms 0.05248387 0.40694539 0.35967366 1.00000000 0.61084492  
## sqft\_living 0.05591911 0.58753189 0.54595425 0.61084492 1.00000000  
## sqft\_lot -0.04881811 0.09879146 0.21379756 -0.15003858 0.21349737  
## floors 0.03142587 0.19599487 -0.01231716 0.48824267 0.21834143  
## waterfront -0.00601867 0.18723046 -0.02388155 0.02402709 0.04512995  
## view 0.04845104 0.32126248 0.04370921 0.10743868 0.20290469  
## condition -0.05442527 0.01080196 0.05075847 -0.20275433 -0.08704616  
## grade 0.06542157 0.62532384 0.17613009 0.51361808 0.62530895  
## sqft\_above 0.06310017 0.47583099 0.37950422 0.52925673 0.80769249  
## sqft\_basement -0.01105019 0.18482790 0.27249273 0.13661508 0.31810283  
## yr\_built 0.07163873 0.11886923 -0.02014115 0.56553240 0.25215473  
## yr\_renovated -0.01141468 0.10260750 0.01761968 0.01698352 0.02780418  
## zipcode -0.03344919 -0.06661544 -0.10383298 -0.08873701 -0.13655332  
## lat -0.01028660 0.39782088 -0.04362736 0.03452169 0.02941721  
## long 0.09490813 0.06021992 0.09157411 0.15541240 0.21179591  
## sqft\_living15 0.07487106 0.53192136 0.34578440 0.44219575 0.73234542  
## sqft\_lot15 -0.05544805 0.10467684 0.16855680 -0.13671006 0.17733937  
## sqft\_lot floors waterfront view condition  
## id -0.04881811 0.03142587 -0.006018670 0.04845104 -0.05442527  
## price 0.09879146 0.19599487 0.187230461 0.32126248 0.01080196  
## bedrooms 0.21379756 -0.01231716 -0.023881549 0.04370921 0.05075847  
## bathrooms -0.15003858 0.48824267 0.024027088 0.10743868 -0.20275433  
## sqft\_living 0.21349737 0.21834143 0.045129953 0.20290469 -0.08704616  
## sqft\_lot 1.00000000 -0.46764775 0.061281460 0.10171693 0.27240654  
## floors -0.46764775 1.00000000 0.020419644 -0.01755120 -0.36800514  
## waterfront 0.06128146 0.02041964 1.000000000 0.34042155 0.01009095  
## view 0.10171693 -0.01755120 0.340421549 1.00000000 0.03451890  
## condition 0.27240654 -0.36800514 0.010090951 0.03451890 1.00000000  
## grade 0.01799467 0.40495046 0.058898054 0.18464079 -0.19021868  
## sqft\_above 0.10812840 0.46766878 0.019014619 0.06247316 -0.19571109  
## sqft\_basement 0.17132458 -0.39896785 0.042393068 0.22760950 0.17395516  
## yr\_built -0.47268086 0.72279773 -0.025271092 -0.08730803 -0.45609638  
## yr\_renovated 0.04836387 -0.03403060 0.050423835 0.06473431 -0.04580786  
## zipcode -0.20648742 0.02202050 0.049153202 0.11697692 -0.09996766  
## lat -0.05925894 0.05026221 -0.007680641 0.01912452 -0.04700562  
## long 0.15246300 0.07677540 -0.010330764 -0.10532559 -0.04109049  
## sqft\_living15 0.26717897 0.16001569 0.063053624 0.22038305 -0.07723497  
## sqft\_lot15 0.80071360 -0.40438723 0.101403098 0.10986952 0.24966404  
## grade sqft\_above sqft\_basement yr\_built  
## id 0.0654215660 0.063100165 -0.0110501872 0.07163873  
## price 0.6253238358 0.475830987 0.1848279021 0.11886923  
## bedrooms 0.1761300924 0.379504224 0.2724927258 -0.02014115  
## bathrooms 0.5136180759 0.529256730 0.1366150794 0.56553240  
## sqft\_living 0.6253089478 0.807692487 0.3181028274 0.25215473  
## sqft\_lot 0.0179946663 0.108128401 0.1713245799 -0.47268086  
## floors 0.4049504567 0.467668775 -0.3989678500 0.72279773  
## waterfront 0.0588980541 0.019014619 0.0423930680 -0.02527109  
## view 0.1846407902 0.062473162 0.2276095048 -0.08730803  
## condition -0.1902186768 -0.195711088 0.1739551648 -0.45609638  
## grade 1.0000000000 0.629292059 -0.0008512015 0.41391909  
## sqft\_above 0.6292920588 1.000000000 -0.3020484503 0.43178138  
## sqft\_basement -0.0008512015 -0.302048450 1.0000000000 -0.28659200  
## yr\_built 0.4139190905 0.431781375 -0.2865920034 1.00000000  
## yr\_renovated 0.0137545251 0.005527641 0.0360665829 -0.14488568  
## zipcode -0.0963141760 -0.207857824 0.1134416369 -0.10071641  
## lat 0.1244168337 -0.030479955 0.0965731536 -0.05718985  
## long 0.1212933902 0.346947946 -0.2154373008 0.25864663  
## sqft\_living15 0.6095840261 0.677956453 0.0939590371 0.21331396  
## sqft\_lot15 0.0242929122 0.088065155 0.1451238319 -0.41247570  
## yr\_renovated zipcode lat long  
## id -0.011414678 -0.03344919 -0.010286601 0.09490813  
## price 0.102607501 -0.06661544 0.397820880 0.06021992  
## bedrooms 0.017619683 -0.10383298 -0.043627362 0.09157411  
## bathrooms 0.016983524 -0.08873701 0.034521688 0.15541240  
## sqft\_living 0.027804177 -0.13655332 0.029417207 0.21179591  
## sqft\_lot 0.048363871 -0.20648742 -0.059258942 0.15246300  
## floors -0.034030600 0.02202050 0.050262205 0.07677540  
## waterfront 0.050423835 0.04915320 -0.007680641 -0.01033076  
## view 0.064734312 0.11697692 0.019124515 -0.10532559  
## condition -0.045807864 -0.09996766 -0.047005621 -0.04109049  
## grade 0.013754525 -0.09631418 0.124416834 0.12129339  
## sqft\_above 0.005527641 -0.20785782 -0.030479955 0.34694795  
## sqft\_basement 0.036066583 0.11344164 0.096573154 -0.21543730  
## yr\_built -0.144885684 -0.10071641 -0.057189850 0.25864663  
## yr\_renovated 1.000000000 0.04008558 0.021904703 -0.03691168  
## zipcode 0.040085583 1.00000000 0.185270334 -0.54354459  
## lat 0.021904703 0.18527033 1.000000000 -0.08224426  
## long -0.036911680 -0.54354459 -0.082244257 1.00000000  
## sqft\_living15 0.002850830 -0.25878003 0.012087716 0.33806073  
## sqft\_lot15 0.060629185 -0.16726839 -0.043272796 0.13205797  
## sqft\_living15 sqft\_lot15  
## id 0.07487106 -0.05544805  
## price 0.53192136 0.10467684  
## bedrooms 0.34578440 0.16855680  
## bathrooms 0.44219575 -0.13671006  
## sqft\_living 0.73234542 0.17733937  
## sqft\_lot 0.26717897 0.80071360  
## floors 0.16001569 -0.40438723  
## waterfront 0.06305362 0.10140310  
## view 0.22038305 0.10986952  
## condition -0.07723497 0.24966404  
## grade 0.60958403 0.02429291  
## sqft\_above 0.67795645 0.08806515  
## sqft\_basement 0.09395904 0.14512383  
## yr\_built 0.21331396 -0.41247570  
## yr\_renovated 0.00285083 0.06062919  
## zipcode -0.25878003 -0.16726839  
## lat 0.01208772 -0.04327280  
## long 0.33806073 0.13205797  
## sqft\_living15 1.00000000 0.25196255  
## sqft\_lot15 0.25196255 1.00000000

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

## 'data.frame': 12315 obs. of 19 variables:  
## $ id : num 7.13e+09 6.41e+09 2.49e+09 1.95e+09 1.32e+09 ...  
## $ price : num 221900 538000 604000 510000 257500 ...  
## $ bedrooms : int 3 3 4 3 3 3 3 3 3 4 ...  
## $ bathrooms : num 1 2.25 3 2 2.25 1.5 1 2.5 1.75 3 ...  
## $ sqft\_living : int 1180 2570 1960 1680 1715 1060 1780 1890 1370 2950 ...  
## $ sqft\_lot : int 5650 7242 5000 8080 6819 9711 7470 6560 9680 5000 ...  
## $ waterfront : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ view : int 0 0 0 0 0 0 0 0 0 3 ...  
## $ condition : int 3 3 5 3 3 3 3 3 4 3 ...  
## $ grade : int 7 7 7 8 7 7 7 7 7 9 ...  
## $ sqft\_above : int 1180 2170 1050 1680 1715 1060 1050 1890 1370 1980 ...  
## $ sqft\_basement: int 0 400 910 0 0 0 730 0 0 970 ...  
## $ yr\_built : int 1955 1951 1965 1987 1995 1963 1960 2003 1977 1979 ...  
## $ yr\_renovated : int 0 1991 0 0 0 0 0 0 0 0 ...  
## $ zipcode : int 98178 98125 98136 98074 98003 98198 98146 98038 98074 98126 ...  
## $ lat : num 47.5 47.7 47.5 47.6 47.3 ...  
## $ long : num -122 -122 -122 -122 -122 ...  
## $ sqft\_living15: int 1340 1690 1360 1800 2238 1650 1780 2390 1370 2140 ...  
## $ sqft\_lot15 : int 5650 7639 5000 7503 6819 9711 8113 7570 10208 4000 ...

colnames(HousePrice)

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

set.seed(1)  
for(i in seq(from=0.6, to=0.9, by=0.05)){  
 rn\_train <- sample(nrow(HousePrice),floor(nrow(HousePrice)\*i))  
 train <- HousePrice[rn\_train,colnames(HousePrice)]  
 test <- HousePrice[-rn\_train,colnames(HousePrice)]  
 lm.price<-lm(price~.,data = train)  
 prediction <- predict(lm.price,newdata = test)  
 training\_data\_prediction = fitted(lm.price)  
 training\_rmse = sqrt(sum((training\_data\_prediction-train$price)^2)/nrow(train))  
 testing\_rmse = sqrt(sum((prediction - test$price)^2)/nrow(test))  
 cat("\ntraining\_rmse:",training\_rmse)  
 cat("\ntesting\_rmse:",testing\_rmse)  
 print(lm.price)  
}

## Warning in predict.lm(lm.price, newdata = test): prediction from a rank-  
## deficient fit may be misleading

##   
## training\_rmse: 139824.5  
## testing\_rmse: 140378  
## Call:  
## lm(formula = price ~ ., data = train)  
##   
## Coefficients:  
## (Intercept) id bedrooms bathrooms sqft\_living   
## 3.889e+06 -7.000e-07 -1.401e+04 2.854e+04 9.516e+01   
## sqft\_lot waterfront view condition grade   
## -3.368e+00 3.870e+05 5.475e+04 3.793e+04 8.987e+04   
## sqft\_above sqft\_basement yr\_built yr\_renovated zipcode   
## 2.153e+01 NA -1.240e+03 5.464e+01 -4.231e+02   
## lat long sqft\_living15 sqft\_lot15   
## 6.045e+05 -8.752e+04 2.958e+01 2.329e-03

## Warning in predict.lm(lm.price, newdata = test): prediction from a rank-  
## deficient fit may be misleading

##   
## training\_rmse: 141639.1  
## testing\_rmse: 136890.6  
## Call:  
## lm(formula = price ~ ., data = train)  
##   
## Coefficients:  
## (Intercept) id bedrooms bathrooms sqft\_living   
## 2.861e+06 -8.394e-07 -1.666e+04 2.259e+04 1.091e+02   
## sqft\_lot waterfront view condition grade   
## -4.735e+00 4.036e+05 5.339e+04 4.191e+04 8.844e+04   
## sqft\_above sqft\_basement yr\_built yr\_renovated zipcode   
## 1.221e+01 NA -1.033e+03 5.806e+01 -4.369e+02   
## lat long sqft\_living15 sqft\_lot15   
## 6.080e+05 -1.024e+05 3.424e+01 5.624e-01

## Warning in predict.lm(lm.price, newdata = test): prediction from a rank-  
## deficient fit may be misleading

##   
## training\_rmse: 138996.3  
## testing\_rmse: 142169.6  
## Call:  
## lm(formula = price ~ ., data = train)  
##   
## Coefficients:  
## (Intercept) id bedrooms bathrooms sqft\_living   
## 7.598e+05 -3.916e-07 -1.513e+04 2.682e+04 1.006e+02   
## sqft\_lot waterfront view condition grade   
## -3.880e+00 4.175e+05 4.966e+04 4.458e+04 9.136e+04   
## sqft\_above sqft\_basement yr\_built yr\_renovated zipcode   
## 1.416e+01 NA -1.185e+03 5.823e+01 -4.135e+02   
## lat long sqft\_living15 sqft\_lot15   
## 6.072e+05 -1.032e+05 3.844e+01 2.287e-01

## Warning in predict.lm(lm.price, newdata = test): prediction from a rank-  
## deficient fit may be misleading

##   
## training\_rmse: 138042.5  
## testing\_rmse: 145763.6  
## Call:  
## lm(formula = price ~ ., data = train)  
##   
## Coefficients:  
## (Intercept) id bedrooms bathrooms sqft\_living   
## 2.064e+06 -6.147e-07 -1.520e+04 2.437e+04 1.051e+02   
## sqft\_lot waterfront view condition grade   
## -4.142e+00 5.012e+05 4.700e+04 4.309e+04 8.986e+04   
## sqft\_above sqft\_basement yr\_built yr\_renovated zipcode   
## 5.460e+00 NA -1.158e+03 6.738e+01 -3.966e+02   
## lat long sqft\_living15 sqft\_lot15   
## 5.968e+05 -8.271e+04 4.329e+01 -9.905e-02

## Warning in predict.lm(lm.price, newdata = test): prediction from a rank-  
## deficient fit may be misleading

##   
## training\_rmse: 139299.7  
## testing\_rmse: 142494.5  
## Call:  
## lm(formula = price ~ ., data = train)  
##   
## Coefficients:  
## (Intercept) id bedrooms bathrooms sqft\_living   
## 1.856e+06 -6.948e-07 -1.378e+04 2.008e+04 1.003e+02   
## sqft\_lot waterfront view condition grade   
## -3.750e+00 4.704e+05 4.871e+04 4.212e+04 9.146e+04   
## sqft\_above sqft\_basement yr\_built yr\_renovated zipcode   
## 1.359e+01 NA -1.057e+03 5.699e+01 -4.187e+02   
## lat long sqft\_living15 sqft\_lot15   
## 5.989e+05 -9.962e+04 3.980e+01 -2.290e-01

## Warning in predict.lm(lm.price, newdata = test): prediction from a rank-  
## deficient fit may be misleading

##   
## training\_rmse: 140995.1  
## testing\_rmse: 133889.9  
## Call:  
## lm(formula = price ~ ., data = train)  
##   
## Coefficients:  
## (Intercept) id bedrooms bathrooms sqft\_living   
## 1.627e+06 -5.047e-07 -1.503e+04 2.355e+04 9.974e+01   
## sqft\_lot waterfront view condition grade   
## -3.847e+00 4.529e+05 5.014e+04 4.229e+04 9.171e+04   
## sqft\_above sqft\_basement yr\_built yr\_renovated zipcode   
## 1.512e+01 NA -1.100e+03 5.384e+01 -4.443e+02   
## lat long sqft\_living15 sqft\_lot15   
## 6.047e+05 -1.204e+05 4.029e+01 2.662e-01

## Warning in predict.lm(lm.price, newdata = test): prediction from a rank-  
## deficient fit may be misleading

##   
## training\_rmse: 139423.3  
## testing\_rmse: 144524.3  
## Call:  
## lm(formula = price ~ ., data = train)  
##   
## Coefficients:  
## (Intercept) id bedrooms bathrooms sqft\_living   
## 1.953e+06 -1.025e-06 -1.468e+04 2.345e+04 1.068e+02   
## sqft\_lot waterfront view condition grade   
## -3.779e+00 4.648e+05 5.086e+04 4.132e+04 9.088e+04   
## sqft\_above sqft\_basement yr\_built yr\_renovated zipcode   
## 9.914e+00 NA -1.106e+03 5.808e+01 -4.205e+02   
## lat long sqft\_living15 sqft\_lot15   
## 6.052e+05 -9.874e+04 3.513e+01 -1.563e-01