# Data Importing And Cleaning

HousePrice <- read.csv("G:/Ryerson-BigData/capstone/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"

any(is.na(HousePrice$id))

## [1] FALSE

any(is.na(HousePrice$date))

## [1] FALSE

any(is.na(HousePrice$price))

## [1] FALSE

any(is.na(HousePrice$bedrooms))

## [1] FALSE

any(is.na(HousePrice$bathrooms))

## [1] FALSE

any(is.na(HousePrice$sqft\_living))

## [1] FALSE

any(is.na(HousePrice$sqft\_lot))

## [1] FALSE

any(is.na(HousePrice$floors))

## [1] FALSE

any(is.na(HousePrice$waterfront))

## [1] FALSE

any(is.na(HousePrice$view))

## [1] FALSE

any(is.na(HousePrice$condition))

## [1] FALSE

any(is.na(HousePrice$grade))

## [1] FALSE

any(is.na(HousePrice$sqft\_above))

## [1] FALSE

any(is.na(HousePrice$sqft\_basement))

## [1] FALSE

any(is.na(HousePrice$yr\_built))

## [1] FALSE

any(is.na(HousePrice$yr\_renovated))

## [1] FALSE

any(is.na(HousePrice$zipcode))

## [1] FALSE

any(is.na(HousePrice$lat))

## [1] FALSE

any(is.na(HousePrice$long))

## [1] FALSE

any(is.na(HousePrice$sqft\_living15))

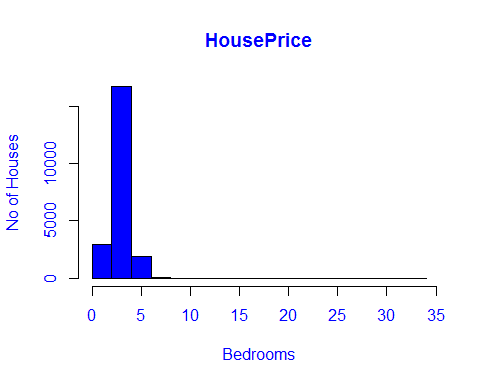
## [1] FALSE

any(is.na(HousePrice$sqft\_lot15))

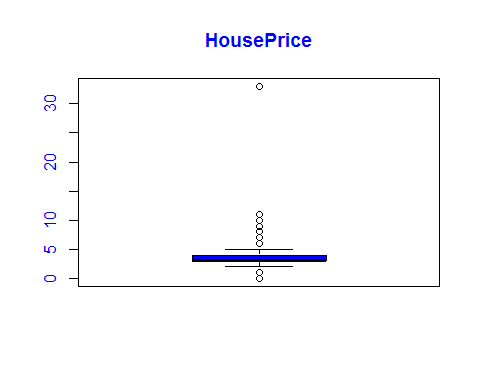
## [1] FALSE

# BEDROOM ANALYSIS

hist(HousePrice$bedrooms,main = 'HousePrice',xlab = 'Bedrooms',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



boxplot(HousePrice$bedrooms,main = 'HousePrice',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



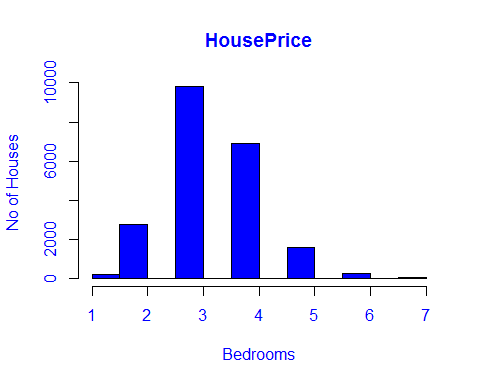
#\*\*\*\*\*\*\*Initial correlation =   
cor(HousePrice$bedrooms,HousePrice$price)

## [1] 0.3083496

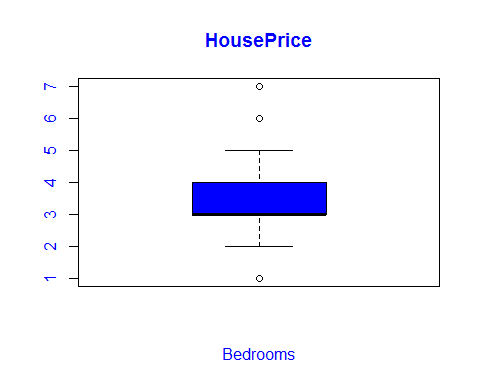
#\*\*\*\*\*\*\*Removing the outliers  
#Since more than 7 bedrooms are very rare.Also it's the outlier for my model.  
#I have removed the outlier data.  
HousePrice<-subset(HousePrice,bedrooms>=1 & bedrooms<=7)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
#\*\*\*\*\*\*Final Correlation =  
cor(HousePrice$bedrooms,HousePrice$price)

## [1] 0.3156734

hist(HousePrice$bedrooms,main = 'HousePrice',xlab = 'Bedrooms',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

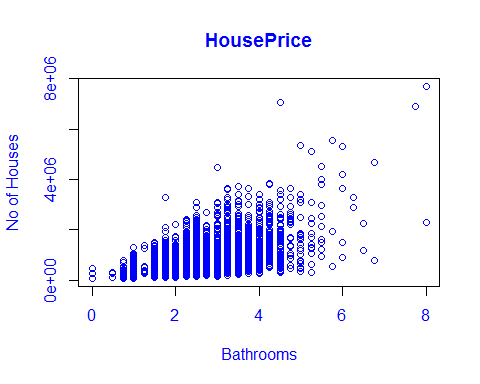


boxplot(HousePrice$bedrooms,main = 'HousePrice',xlab = 'Bedrooms',  
 col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

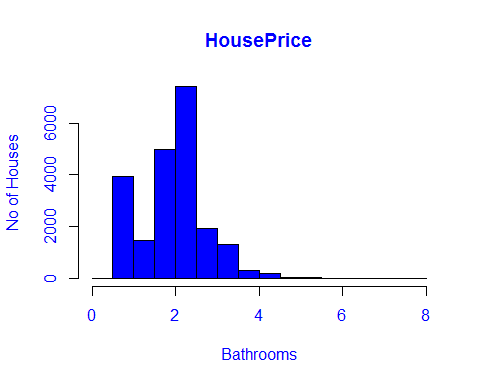


# BATHROOM ANALYSIS

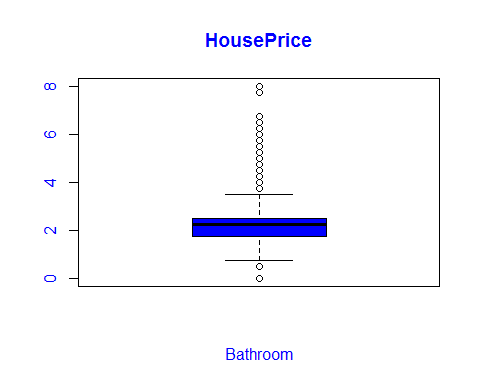
plot(HousePrice$price~HousePrice$bathrooms,main = 'HousePrice',xlab = 'Bathrooms',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



hist(HousePrice$bathrooms,main = 'HousePrice',xlab = 'Bathrooms',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



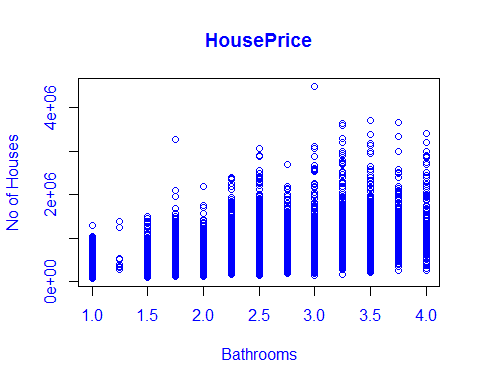
boxplot(HousePrice$bathrooms,main = 'HousePrice',xlab = 'Bathroom',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



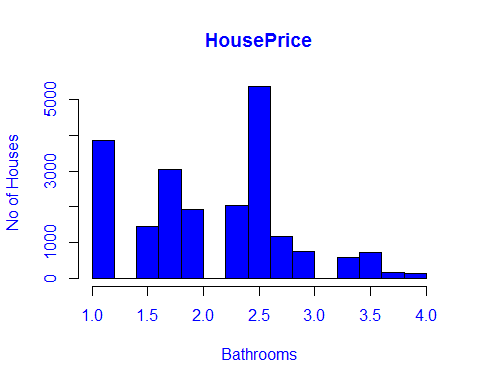
#\*\*\*\*\*\*\*Initial correlation =   
cor(HousePrice$bathrooms,HousePrice$price)

## [1] 0.5259342

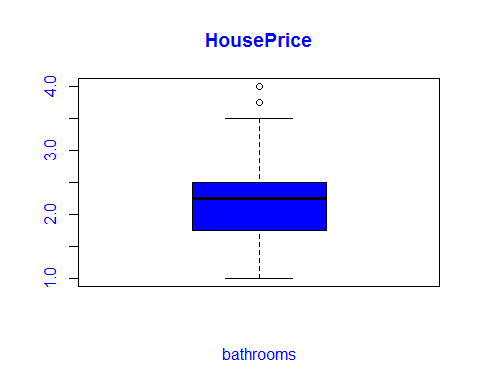
#\*\*\*\*\*\*\*Removing the outliers  
#More than 4 bathrooms are very rare in this data.So I am removing it.  
HousePrice<-subset(HousePrice,bathrooms>=1 & bathrooms<=4)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
plot(HousePrice$price~HousePrice$bathrooms,main = 'HousePrice',xlab = 'Bathrooms',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



hist(HousePrice$bathrooms,main = 'HousePrice',xlab = 'Bathrooms',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



boxplot(HousePrice$bathrooms,main = 'HousePrice',xlab = 'bathrooms',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

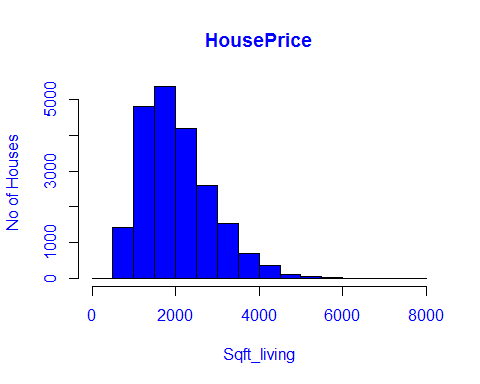


#\*\*\*\*\*\*Final Correlation =  
cor(HousePrice$bathrooms,HousePrice$price)

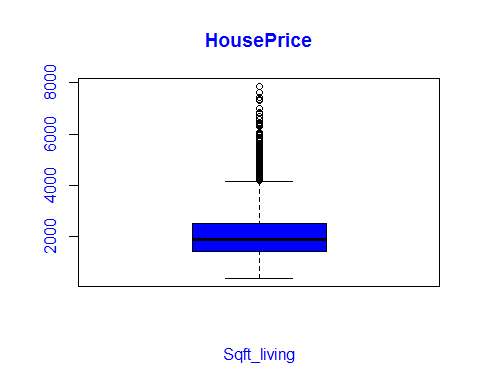
## [1] 0.475159

# SQFT LIVING ANALYSIS

hist(HousePrice$sqft\_living,main = 'HousePrice',xlab = 'Sqft\_living',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



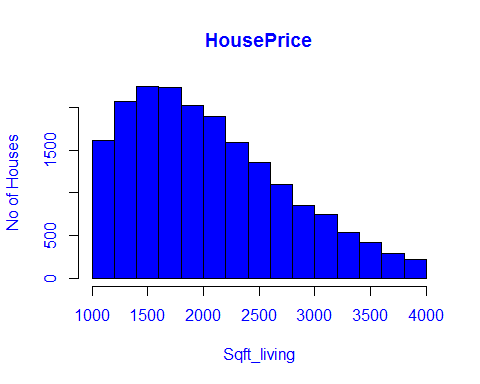
boxplot(HousePrice$sqft\_living, xlab = 'Sqft\_living',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue',main = 'HousePrice')



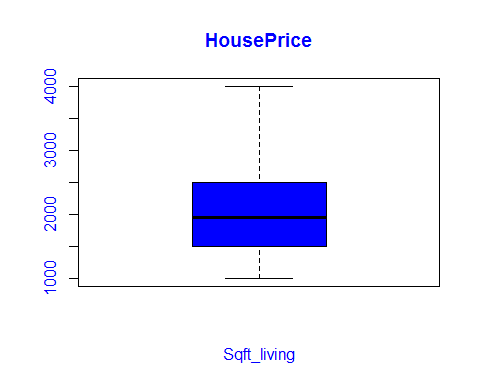
#\*\*\*\*\*\*\*Initial correlation =   
cor(HousePrice$sqft\_living,HousePrice$price)

## [1] 0.6701029

#  
  
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_living >1000 & sqft\_living<=4000)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
hist(HousePrice$sqft\_living,main = 'HousePrice',xlab = 'Sqft\_living',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



boxplot(HousePrice$sqft\_living, xlab = 'Sqft\_living',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue',main = 'HousePrice')

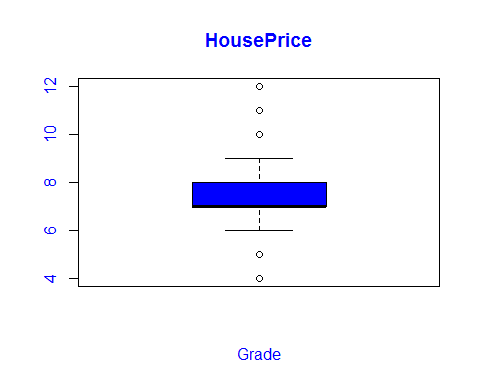


#\*\*\*\*\*\*Final Correlation =  
cor(HousePrice$sqft\_living,HousePrice$price)

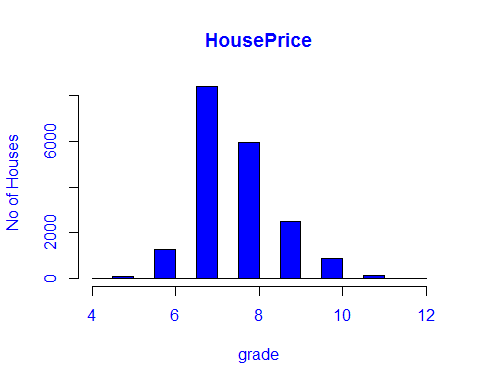
## [1] 0.5938015

## GRADE ANALYSIS

boxplot(HousePrice$grade,main = 'HousePrice',xlab = 'Grade',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



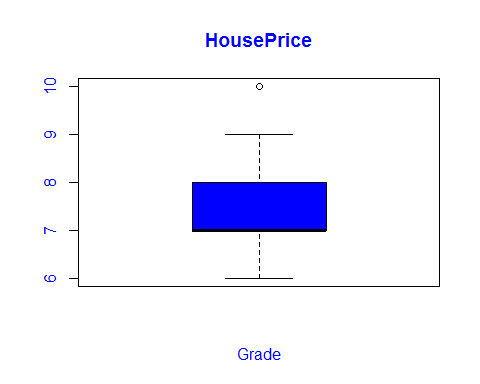
hist(HousePrice$grade,main = 'HousePrice',xlab = 'grade',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



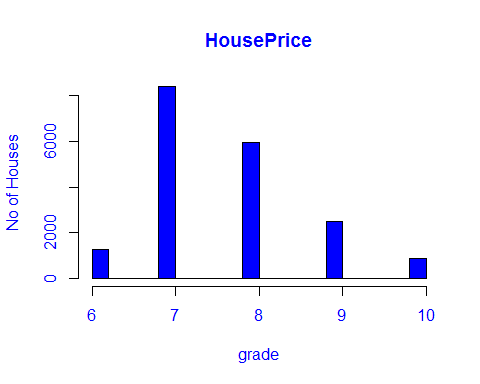
#\*\*\*\*\*\*\*Initial correlation =   
cor(HousePrice$price,HousePrice$grade)

## [1] 0.6106929

#\*\*\*\*\*\*\*Removing the outliers  
  
#Most of the houses grades are between 6-10   
HousePrice<-subset(HousePrice,grade >= 6 & grade<=10)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
  
boxplot(HousePrice$grade,main = 'HousePrice',xlab = 'Grade',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



hist(HousePrice$grade,main = 'HousePrice',xlab = 'grade',ylab = 'No of Houses'  
,col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

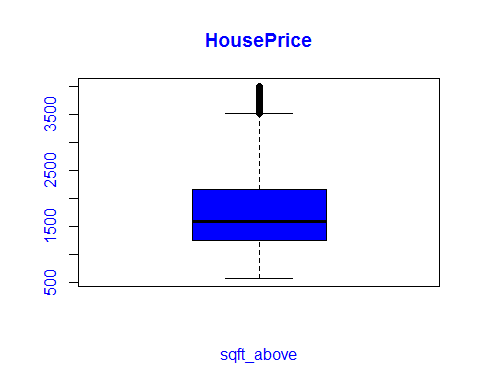


#\*\*\*\*\*\*Final Correlation =   
cor(HousePrice$price,HousePrice$grade)

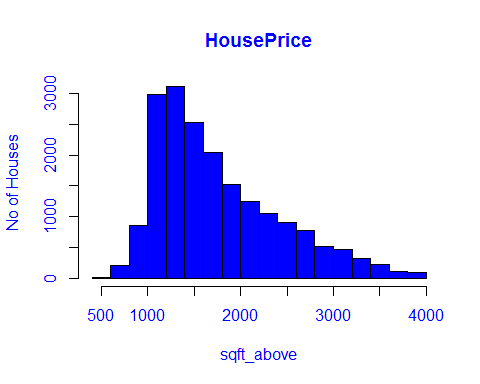
## [1] 0.592697

# SQFT\_ABOVE ANALYSIS

boxplot(HousePrice$sqft\_above,main = 'HousePrice',xlab = 'sqft\_above',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



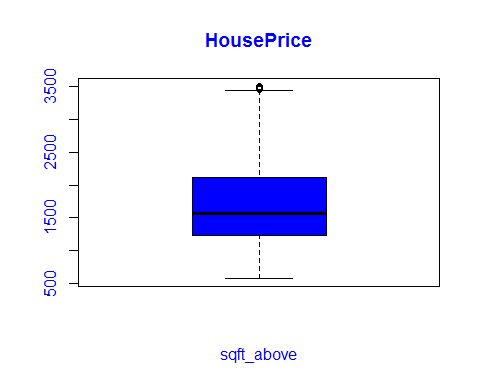
hist(HousePrice$sqft\_above,main = 'HousePrice',xlab = 'sqft\_above',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



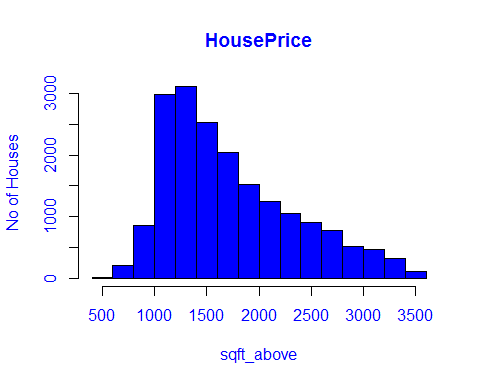
#\*\*\*\*\*\*\*Initial correlation =  
cor(HousePrice$price,HousePrice$sqft\_above)

## [1] 0.4553313

#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_above >=500 & sqft\_above<=3500)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
  
boxplot(HousePrice$sqft\_above,main = 'HousePrice',xlab = 'sqft\_above',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



hist(HousePrice$sqft\_above,main = 'HousePrice',xlab = 'sqft\_above',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

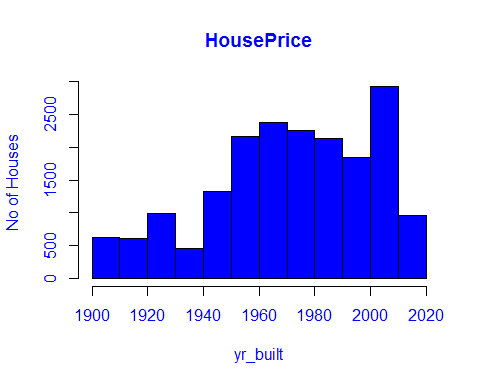


#\*\*\*\*\*\*Final Correlation =   
cor(HousePrice$price,HousePrice$sqft\_above)

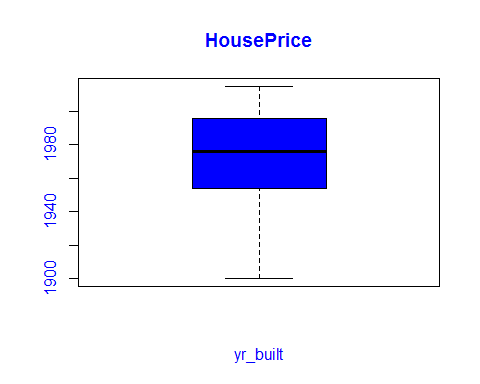
## [1] 0.4254812

# YR\_BUILT ANALYSIS

##YR\_BUILT ANALYSIS  
hist(HousePrice$yr\_built,main = 'HousePrice',xlab = 'yr\_built',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



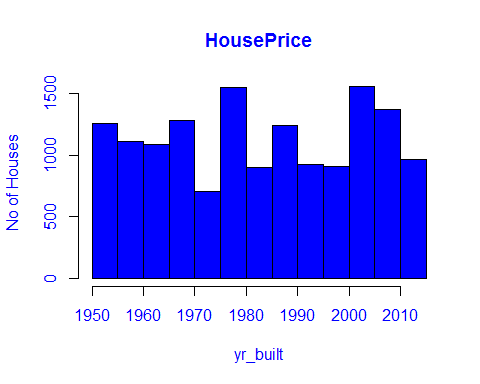
boxplot(HousePrice$yr\_built,main = 'HousePrice',xlab = 'yr\_built',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



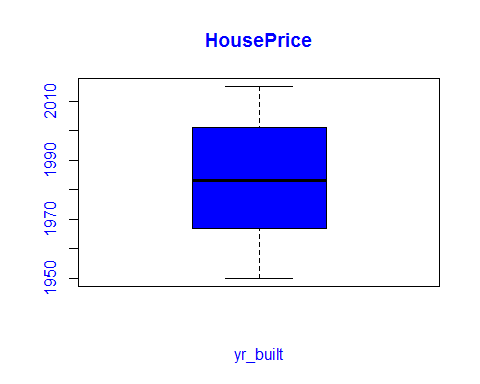
#\*\*\*\*\*\*\*Initial correlation =  
cor(HousePrice$price,HousePrice$yr\_built)

## [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)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
  
hist(HousePrice$yr\_built,main = 'HousePrice',xlab = 'yr\_built',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



boxplot(HousePrice$yr\_built,main = 'HousePrice',xlab = 'yr\_built',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

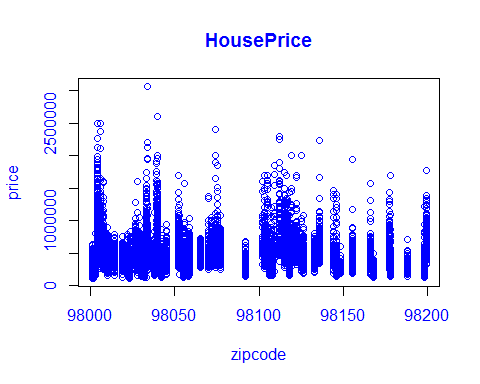


#\*\*\*\*\*\*Final Correlation =   
cor(HousePrice$price,HousePrice$yr\_built)

## [1] 0.08555947

# ZIPCODE ANALYSIS

plot(HousePrice$price~HousePrice$zipcode,main = 'HousePrice',xlab = 'zipcode',ylab = 'price',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

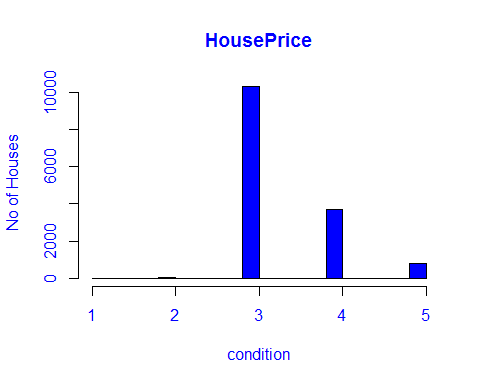


#correlation =   
cor(HousePrice$price,HousePrice$zipcode)

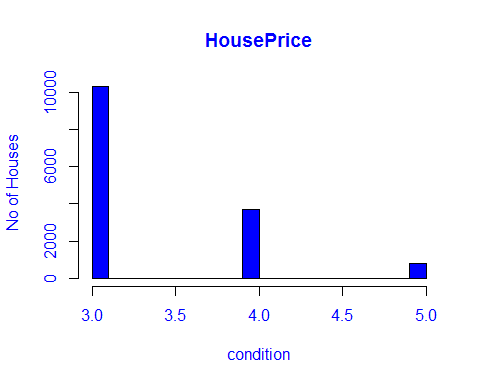
## [1] -0.02647096

## CONDITION aNALYSIS

hist(HousePrice$condition,main = 'HousePrice',xlab = 'condition',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



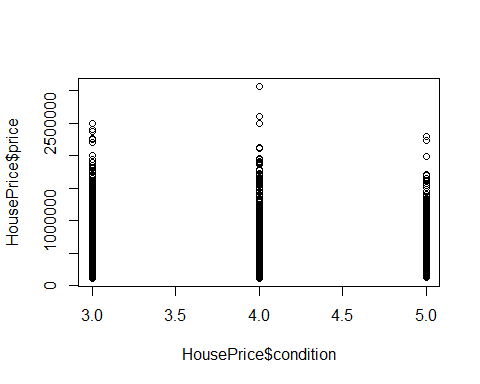
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,condition>=3& condition<=5)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
# Most of the houses are of condition 3-5  
  
hist(HousePrice$condition,main = 'HousePrice',xlab = 'condition',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



#\*\*\*\*\*\*\*\*\*correlation =   
cor(HousePrice$price,HousePrice$condition)

## [1] 0.02442002

plot(HousePrice$price~HousePrice$condition)

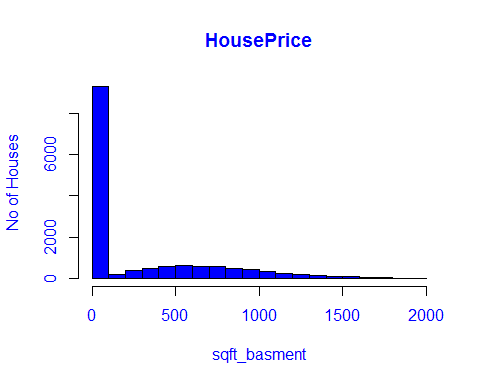


## SQFT\_BASEMENT ANALYSIS

#\*\*\*\*\*\*\*Initial correlation =   
cor(HousePrice$price,HousePrice$sqft\_basement)

## [1] 0.2254519

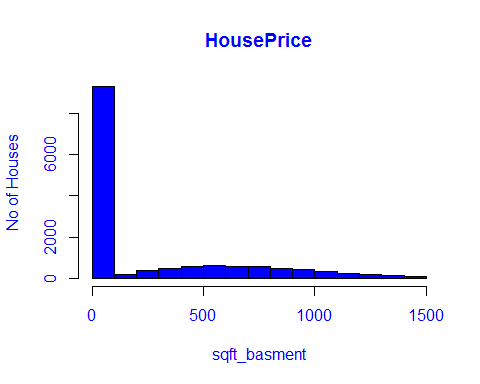
hist(HousePrice$sqft\_basement,main = 'HousePrice',xlab = 'sqft\_basment',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



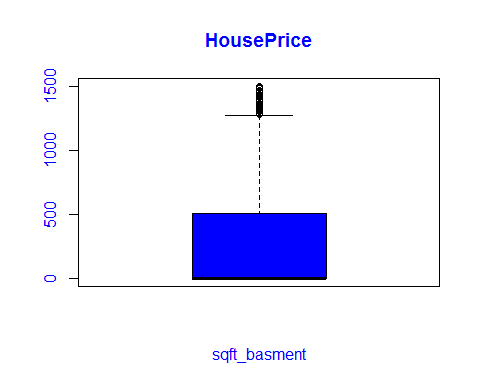
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_basement >=0 & sqft\_basement<=1500)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
#\*\*\*\*\*\*\*\*\*\*\*\*\*Final correlation =   
cor(HousePrice$price,HousePrice$sqft\_basement)

## [1] 0.1879221

hist(HousePrice$sqft\_basement,main = 'HousePrice',xlab = 'sqft\_basment',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

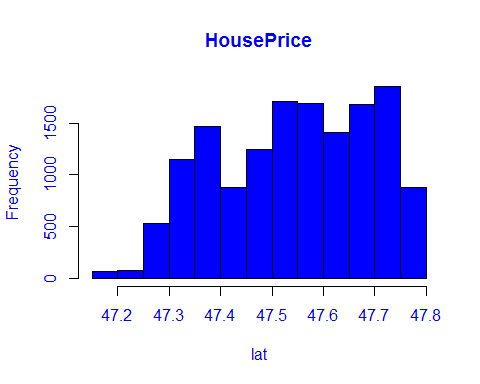


boxplot(HousePrice$sqft\_basement,main = 'HousePrice',xlab = 'sqft\_basment',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

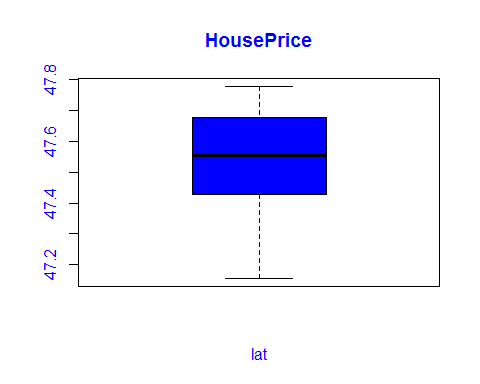


## LAT ANALYSIS

hist(HousePrice$lat, main = 'HousePrice',xlab = 'lat',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



boxplot(HousePrice$lat,main = 'HousePrice',xlab = 'lat',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



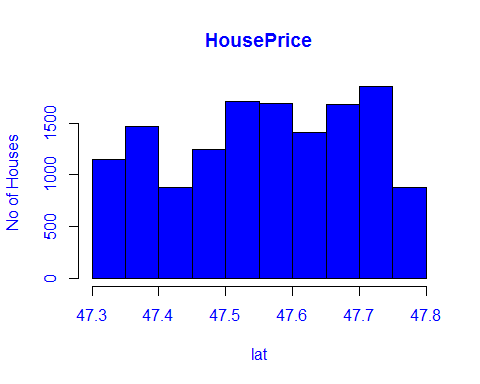
#\*\*\*\*\*\*\*\*\*Initial Correlation =  
cor(HousePrice$price,HousePrice$lat)

## [1] 0.4161934

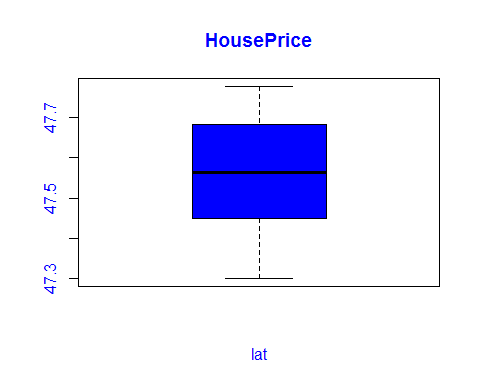
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,lat>=47.3)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
#\*\*\*\*\*\*\*\*\*\*\*\*Final Correlation =   
cor(HousePrice$price,HousePrice$lat)

## [1] 0.3848454

hist(HousePrice$lat, main = 'HousePrice',xlab = 'lat',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

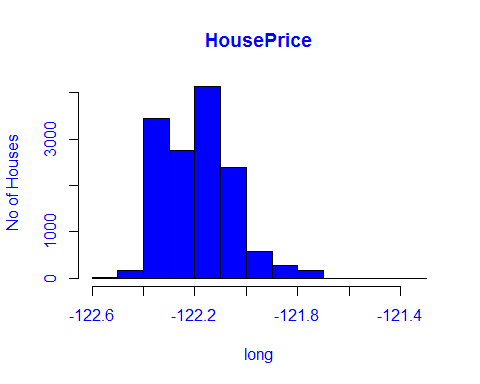


boxplot(HousePrice$lat,main = 'HousePrice',xlab = 'lat',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



## LONG ANALYSIS

hist(HousePrice$long,main = 'HousePrice',xlab = 'long',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



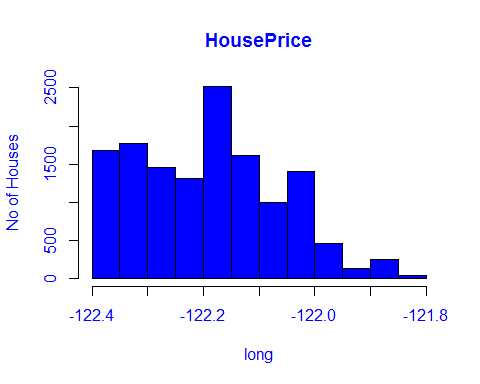
#\*\*\*\*\*\*\*\*\*Initial Correlation =  
cor(HousePrice$price,HousePrice$long)

## [1] 0.03740839

#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,long>=-122.4 & long < -121.8)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
#\*\*\*\*\*\*\*\*\*\*\*\*Final Correlation =   
cor(HousePrice$price,HousePrice$long)

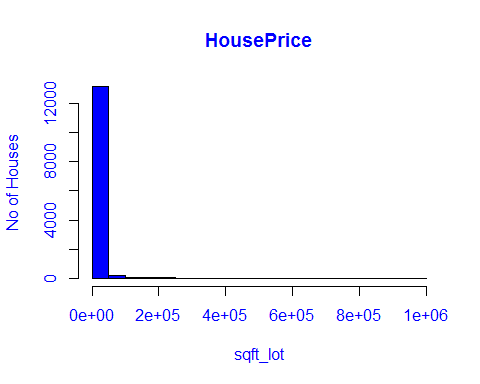
## [1] 0.0746599

hist(HousePrice$long,main = 'HousePrice',xlab = 'long',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

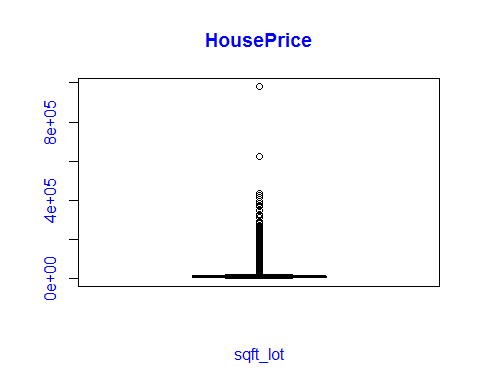


## SQFT\_LOT ANALYSIS

hist(HousePrice$sqft\_lot,main = 'HousePrice',xlab = 'sqft\_lot',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



boxplot(HousePrice$sqft\_lot,main = 'HousePrice',xlab = 'sqft\_lot', col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



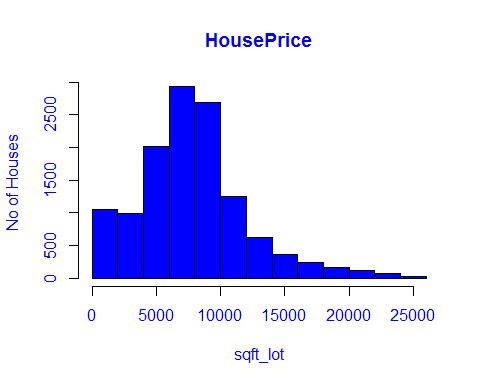
#\*\*\*\*\*\*\*\*\*Initial Correlation =  
cor(HousePrice$price,HousePrice$sqft\_lot)

## [1] 0.09390182

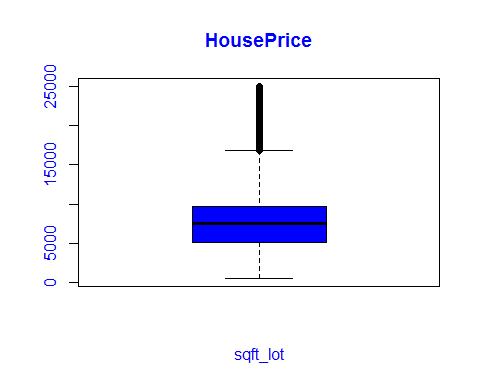
#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_lot>=0 & sqft\_lot<=25000)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
#\*\*\*\*\*\*\*\*\*\*\*\*Final Correlation =   
cor(HousePrice$price,HousePrice$sqft\_lot)

## [1] 0.1228599

hist(HousePrice$sqft\_lot,main = 'HousePrice',xlab = 'sqft\_lot',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

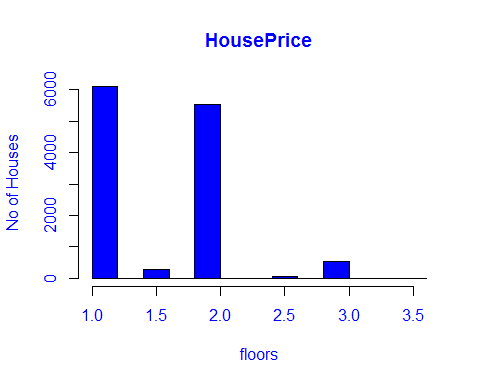


boxplot(HousePrice$sqft\_lot,main = 'HousePrice',xlab = 'sqft\_lot', col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

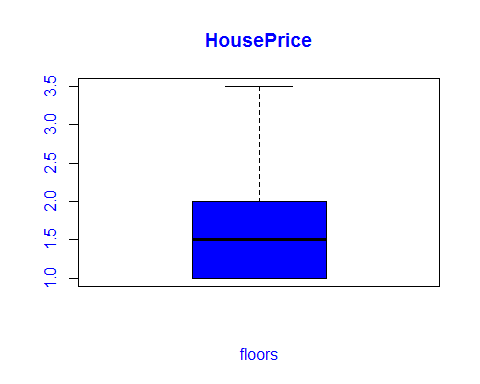


## FLOOR ANALYSIS

hist(HousePrice$floors,main = 'HousePrice',xlab = 'floors',ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



boxplot(HousePrice$floors,main = 'HousePrice',xlab = 'floors' ,col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

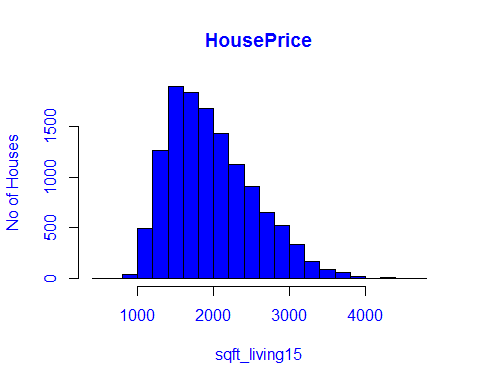


#\*\*\*\*\*\*\*\*\*Initial Correlation =  
cor(HousePrice$price,HousePrice$floors)

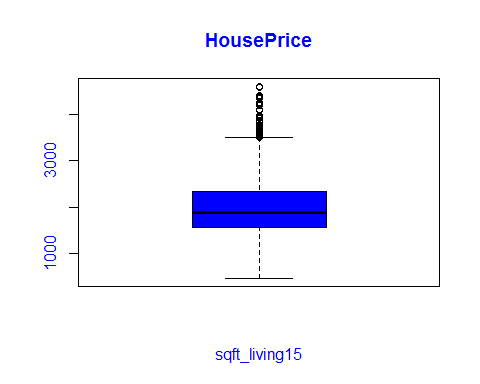
## [1] 0.189869

## SQFT\_LIVING15 ANALYSIS

hist(HousePrice$sqft\_living15, main = 'HousePrice',xlab = 'sqft\_living15', ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue', col = 'Blue')



boxplot(HousePrice$sqft\_living15,main = 'HousePrice',xlab = 'sqft\_living15', col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')

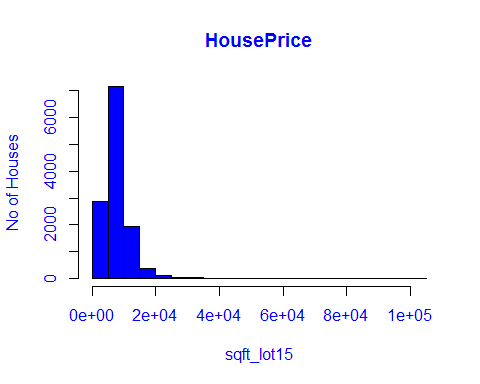


#\*\*\*\*\*\*\*\*\*\*\*\*Final Correlation =   
cor(HousePrice$sqft\_living15,HousePrice$price )

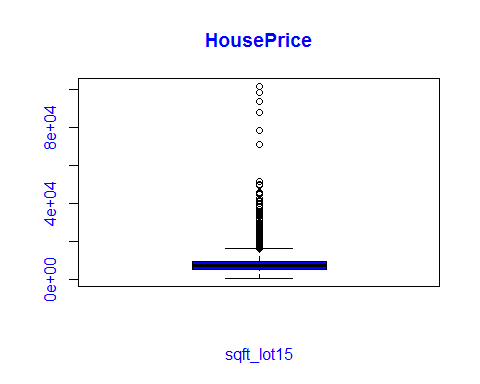
## [1] 0.5377631

## SQFT\_LOT15 ANALYSIS

hist(HousePrice$sqft\_lot15, main = 'HousePrice',xlab = 'sqft\_lot15', ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue', col = 'Blue')



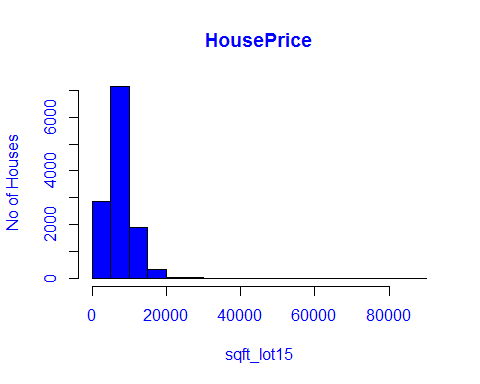
boxplot(HousePrice$sqft\_lot15,main = 'HousePrice',xlab = 'sqft\_lot15',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



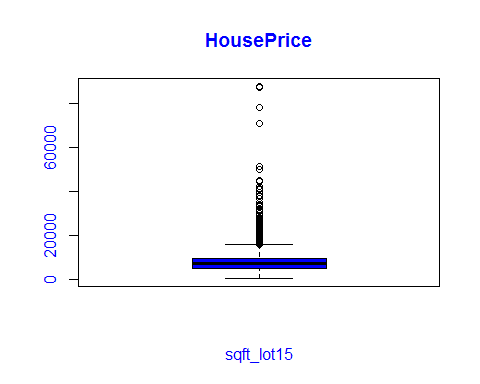
#\*\*\*\*\*\*\*\*\*\*\*\*initial Correlation =   
cor(HousePrice$sqft\_lot15,HousePrice$price)

## [1] 0.1155467

#\*\*\*\*\*\*\*Removing the outliers  
HousePrice<-subset(HousePrice,sqft\_lot15>=0 & sqft\_lot<=20000)  
  
#\*\*\*\*\*\*\*Once we removed the outliers  
hist(HousePrice$sqft\_lot15, main = 'HousePrice',xlab = 'sqft\_lot15', ylab = 'No of Houses',col.main='Blue',col.axis='Blue',col.lab = 'Blue', col = 'Blue')



boxplot(HousePrice$sqft\_lot15,main = 'HousePrice',xlab = 'sqft\_lot15',col.main='Blue',col.axis='Blue',col.lab = 'Blue',col = 'Blue')



#\*\*\*\*\*\*\*\*\*\*\*\*Final Correlation =   
cor(HousePrice$sqft\_lot15,HousePrice$price)

## [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

# USING BORUTA for features selection and finally making model with the selected attributes

install.packages("Boruta", repos = "http://cran.us.r-project.org")

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

## Warning: unable to access index for repository http://cran.us.r-project.org/src/contrib:  
## cannot open URL 'http://cran.us.r-project.org/src/contrib/PACKAGES'

## Warning: package 'Boruta' is not available (for R version 3.3.0)

## Warning: unable to access index for repository http://cran.us.r-project.org/bin/windows/contrib/3.3:  
## cannot open URL 'http://cran.us.r-project.org/bin/windows/contrib/3.3/PACKAGES'

library(Boruta)

## Warning: package 'Boruta' was built under R version 3.3.2

## Loading required package: ranger

## Warning: package 'ranger' was built under R version 3.3.2

set.seed(2)  
boruta.train <- Boruta(price ~., data = HousePrice , doTrace = 2,ntree = 500)

## 1. run of importance source...

## 2. run of importance source...

## 3. run of importance source...

## 4. run of importance source...

## 5. run of importance source...

## 6. run of importance source...

## 7. run of importance source...

## 8. run of importance source...

## 9. run of importance source...

## 10. run of importance source...

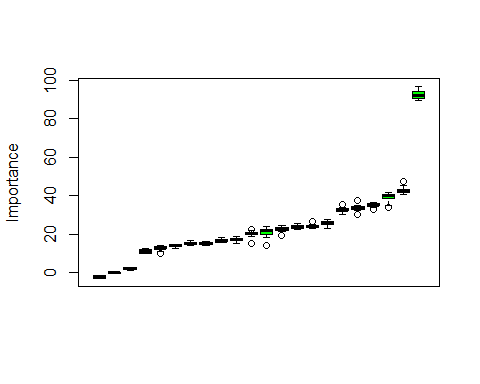
## 11. run of importance source...

## Confirmed 19 attributes: bathrooms, bedrooms, condition, floors, grade and 14 more.

print(boruta.train)

## Boruta performed 11 iterations in 4.41847 mins.  
## 19 attributes confirmed important: bathrooms, bedrooms,  
## condition, floors, grade and 14 more.  
## No attributes deemed unimportant.

plot(boruta.train, xlab = "", xaxt = "n")



set.seed(1)  
Boruta.Short <- Boruta(price ~ ., data = HousePrice, maxRuns = 12)

HousePrice$date <- NULL  
rn\_train <- sample(nrow(HousePrice),floor(nrow(HousePrice)\*0.6))  
train <- HousePrice[rn\_train,colnames(HousePrice)]  
test <- HousePrice[-rn\_train,colnames(HousePrice)]  
lm <- lm(formula = price~.,data=train)  
prediction <- predict(lm,newdata = test)

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

training\_data\_prediction = fitted(lm)  
training\_rmse = sqrt(sum((training\_data\_prediction-train$price)^2)/nrow(train))  
testing\_rmse = sqrt(sum((prediction - test$price)^2)/nrow(test))  
training\_rmse

## [1] 138170.7

testing\_rmse

## [1] 142802.7

lm

##   
## Call:  
## lm(formula = price ~ ., data = train)  
##   
## Coefficients:  
## (Intercept) id bedrooms bathrooms sqft\_living   
## 1.255e+06 -8.985e-07 -1.423e+04 2.313e+04 1.084e+02   
## sqft\_lot floors waterfront view condition   
## -4.369e+00 4.884e+03 5.415e+05 4.456e+04 4.194e+04   
## grade sqft\_above sqft\_basement yr\_built yr\_renovated   
## 9.034e+04 4.292e+00 NA -1.242e+03 6.062e+01   
## zipcode lat long sqft\_living15 sqft\_lot15   
## -3.946e+02 6.058e+05 -8.550e+04 3.870e+01 2.760e-01