R Notebook

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

House = read.csv("Housing.csv",header = TRUE)

head(House)

## price area bedrooms bathrooms stories mainroad guestroom basement  
## 1 13300000 7420 4 2 3 yes no no  
## 2 12250000 8960 4 4 4 yes no no  
## 3 12250000 9960 3 2 2 yes no yes  
## 4 12215000 7500 4 2 2 yes no yes  
## 5 11410000 7420 4 1 2 yes yes yes  
## 6 10850000 7500 3 3 1 yes no yes  
## hotwaterheating airconditioning parking prefarea furnishingstatus  
## 1 no yes 2 yes furnished  
## 2 no yes 3 no furnished  
## 3 no no 2 yes semi-furnished  
## 4 no yes 3 yes furnished  
## 5 no yes 2 no furnished  
## 6 no yes 2 yes semi-furnished

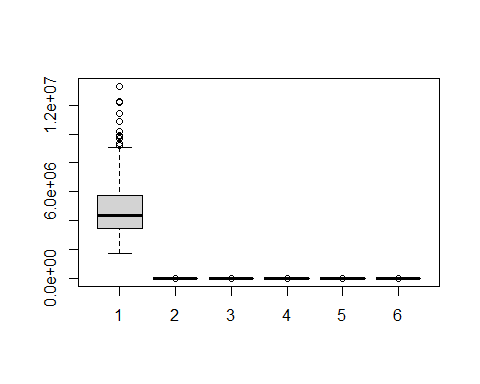
str(House)

## 'data.frame': 545 obs. of 13 variables:  
## $ price : int 13300000 12250000 12250000 12215000 11410000 10850000 10150000 10150000 9870000 9800000 ...  
## $ area : int 7420 8960 9960 7500 7420 7500 8580 16200 8100 5750 ...  
## $ bedrooms : int 4 4 3 4 4 3 4 5 4 3 ...  
## $ bathrooms : int 2 4 2 2 1 3 3 3 1 2 ...  
## $ stories : int 3 4 2 2 2 1 4 2 2 4 ...  
## $ mainroad : chr "yes" "yes" "yes" "yes" ...  
## $ guestroom : chr "no" "no" "no" "no" ...  
## $ basement : chr "no" "no" "yes" "yes" ...  
## $ hotwaterheating : chr "no" "no" "no" "no" ...  
## $ airconditioning : chr "yes" "yes" "no" "yes" ...  
## $ parking : int 2 3 2 3 2 2 2 0 2 1 ...  
## $ prefarea : chr "yes" "no" "yes" "yes" ...  
## $ furnishingstatus: chr "furnished" "furnished" "semi-furnished" "furnished" ...

# Analyzing data-19BDS0144  
summary(House)

## price area bedrooms bathrooms   
## Min. : 1750000 Min. : 1650 Min. :1.000 Min. :1.000   
## 1st Qu.: 3430000 1st Qu.: 3600 1st Qu.:2.000 1st Qu.:1.000   
## Median : 4340000 Median : 4600 Median :3.000 Median :1.000   
## Mean : 4766729 Mean : 5151 Mean :2.965 Mean :1.286   
## 3rd Qu.: 5740000 3rd Qu.: 6360 3rd Qu.:3.000 3rd Qu.:2.000   
## Max. :13300000 Max. :16200 Max. :6.000 Max. :4.000   
## stories mainroad guestroom basement   
## Min. :1.000 Length:545 Length:545 Length:545   
## 1st Qu.:1.000 Class :character Class :character Class :character   
## Median :2.000 Mode :character Mode :character Mode :character   
## Mean :1.806   
## 3rd Qu.:2.000   
## Max. :4.000   
## hotwaterheating airconditioning parking prefarea   
## Length:545 Length:545 Min. :0.0000 Length:545   
## Class :character Class :character 1st Qu.:0.0000 Class :character   
## Mode :character Mode :character Median :0.0000 Mode :character   
## Mean :0.6936   
## 3rd Qu.:1.0000   
## Max. :3.0000   
## furnishingstatus   
## Length:545   
## Class :character   
## Mode :character   
##   
##   
##

#Visualising numerical data for identifying outliers  
boxplot(list(House$price,House$area,House$bedrooms,House$bathrooms,House$stories,House$parking), horizontal=FALSE,outline = TRUE)



#Checking for outliers  
out\_area<- boxplot.stats(House$area)$out  
out\_area

## [1] 16200 13200 11440 11175 13200 12090 15600 11460 11410 10700 12900 12944

out\_price <- boxplot.stats(House$price)$out  
out\_price

## [1] 13300000 12250000 12250000 12215000 11410000 10850000 10150000 10150000  
## [9] 9870000 9800000 9800000 9681000 9310000 9240000 9240000

out\_park<- boxplot.stats(House$parking)$out  
out\_park

## [1] 3 3 3 3 3 3 3 3 3 3 3 3

out\_story<- boxplot.stats(House$stories)$out  
out\_story

## [1] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4  
## [39] 4 4 4

out\_bed<- boxplot.stats(House$bedrooms)$out  
out\_bed

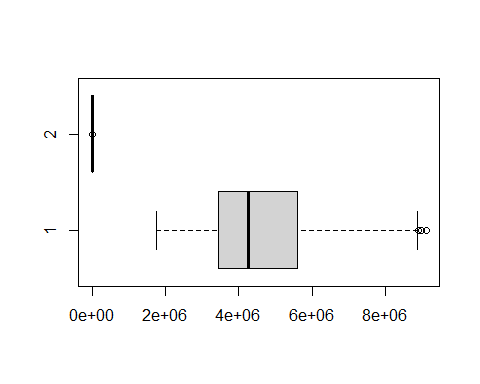
## [1] 5 5 5 5 6 5 5 5 5 5 6 5

out\_bath<- boxplot.stats(House$bathrooms)$out  
out\_bath

## [1] 4

#Price- maximum outliers , Area,bedrooms,stories - next , Bathroom-minimum outliers

# Since we have many observations , we can remove the outliers from price and area as they seem less significant  
#Removing outliers for price and area columns  
df<-House  
df<- df[-which(df$price %in% out\_price),-which(df$area %in% out\_area)]  
boxplot(list(df$price,df$area),horizontal = TRUE)



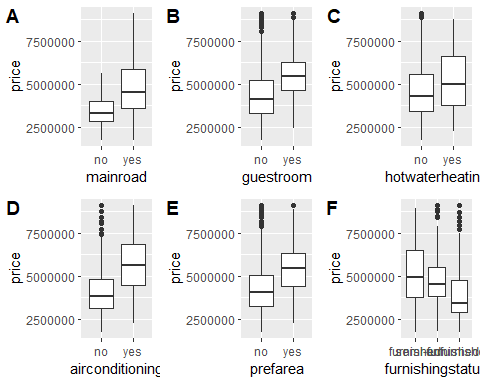
#Visualising categorical data  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.0.5

pm<-ggplot(df, aes(x = price, y = mainroad)) + geom\_boxplot() + coord\_flip()  
  
pg<-ggplot(df, aes(x = price, y = guestroom)) + geom\_boxplot() + coord\_flip()  
  
phot<-ggplot(df, aes(x = price, y = hotwaterheating)) + geom\_boxplot() + coord\_flip()  
  
pac<-ggplot(df, aes(x = price, y = airconditioning)) + geom\_boxplot() + coord\_flip()  
p\_pref<-ggplot(df, aes(x = price, y = prefarea)) + geom\_boxplot() + coord\_flip()  
  
pfurn<- ggplot(df, aes(x = price, y = furnishingstatus)) + geom\_boxplot() + coord\_flip()  
  
library("cowplot")

## Warning: package 'cowplot' was built under R version 4.0.5

plot\_grid(pm, pg,phot,pac,p\_pref,pfurn, labels ="AUTO",  
 ncol = 3, nrow = 2)



# Data preparation - Converting categorical to numerical 19BDS0144  
df\_1<-House  
df\_1$airconditioning<-as.integer(as.factor(df\_1$airconditioning))  
df\_1$mainroad<-as.integer(as.factor(df\_1$mainroad))  
df\_1$hotwaterheating<-as.integer(as.factor(df\_1$hotwaterheating))  
df\_1$prefarea<-as.integer(as.factor(df\_1$prefarea))  
df\_1$guestroom<-as.integer(as.factor(df\_1$guestroom))  
df\_1$furnishingstatus<- as.integer(as.factor(df\_1$furnishingstatus))  
df\_1$basement<-as.integer(as.factor(df\_1$basement))  
str(df\_1)

## 'data.frame': 545 obs. of 13 variables:  
## $ price : int 13300000 12250000 12250000 12215000 11410000 10850000 10150000 10150000 9870000 9800000 ...  
## $ area : int 7420 8960 9960 7500 7420 7500 8580 16200 8100 5750 ...  
## $ bedrooms : int 4 4 3 4 4 3 4 5 4 3 ...  
## $ bathrooms : int 2 4 2 2 1 3 3 3 1 2 ...  
## $ stories : int 3 4 2 2 2 1 4 2 2 4 ...  
## $ mainroad : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ guestroom : int 1 1 1 1 2 1 1 1 2 2 ...  
## $ basement : int 1 1 2 2 2 2 1 1 2 1 ...  
## $ hotwaterheating : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ airconditioning : int 2 2 1 2 2 2 2 1 2 2 ...  
## $ parking : int 2 3 2 3 2 2 2 0 2 1 ...  
## $ prefarea : int 2 1 2 2 1 2 2 1 2 2 ...  
## $ furnishingstatus: int 1 1 2 1 1 2 2 3 1 3 ...

set.seed(1)  
# we will partition into 70% training and 30% validation  
trainrows <- sample(rownames(df\_1), dim(df\_1)[1]\*0.7)  
traindata <- df\_1[trainrows, ]  
# we will set the difference of the training into the validation set i.e. 30%  
validrows <- setdiff(rownames(df\_1), trainrows)  
validdata <- df\_1[validrows, ]

model<-lm(formula= price~area+ bedrooms+bathrooms+stories+mainroad+guestroom+hotwaterheating+airconditioning+prefarea+furnishingstatus ,data= df\_1,subset = trainrows)  
summary(model)

##   
## Call:  
## lm(formula = price ~ area + bedrooms + bathrooms + stories +   
## mainroad + guestroom + hotwaterheating + airconditioning +   
## prefarea + furnishingstatus, data = df\_1, subset = trainrows)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2733372 -664407 -91108 492471 5225016   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.009e+06 6.191e+05 -4.861 1.73e-06 \*\*\*  
## area 2.615e+02 2.957e+01 8.842 < 2e-16 \*\*\*  
## bedrooms 1.059e+05 8.808e+04 1.202 0.230080   
## bathrooms 1.063e+06 1.378e+05 7.710 1.17e-13 \*\*\*  
## stories 3.870e+05 7.674e+04 5.044 7.17e-07 \*\*\*  
## mainroad 5.421e+05 1.856e+05 2.920 0.003713 \*\*   
## guestroom 3.946e+05 1.519e+05 2.598 0.009765 \*\*   
## hotwaterheating 1.063e+06 2.918e+05 3.643 0.000308 \*\*\*  
## airconditioning 8.377e+05 1.361e+05 6.157 1.94e-09 \*\*\*  
## prefarea 7.419e+05 1.417e+05 5.238 2.74e-07 \*\*\*  
## furnishingstatus -2.666e+05 7.894e+04 -3.377 0.000810 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1121000 on 370 degrees of freedom  
## Multiple R-squared: 0.633, Adjusted R-squared: 0.6231   
## F-statistic: 63.82 on 10 and 370 DF, p-value: < 2.2e-16

#For without removing outliers  
#From t values listed below, all the feature variables are significant   
#R-squared indicates 65% of the data fits the regression model

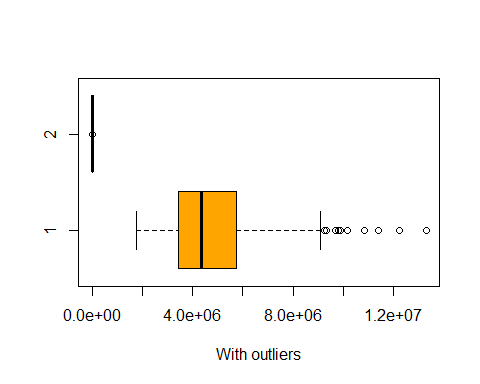
pred<-predict(model,newdata= validdata)  
vres<-data.frame(validdata$price,pred,residuals=validdata$price-pred)  
head(vres)#We can see that even though the model had moderate R squared on the training data the model does not fit quite well on the validation data, as many of the initial values for the Sale price prediction vs actual have quite difference.

## validdata.price pred residuals  
## 5 11410000 6278024 5131976.4  
## 6 10850000 8011952 2838048.1  
## 7 10150000 9561317 588683.3  
## 8 10150000 9039297 1110703.2  
## 9 9870000 7197739 2672261.1  
## 10 9800000 7780867 2019132.5

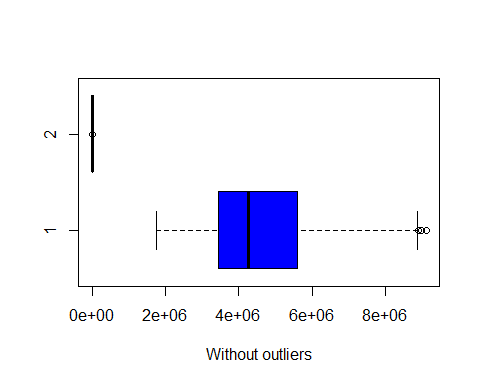
#Confidence interval  
confint(model)

## 2.5 % 97.5 %  
## (Intercept) -4226731.8005 -1791840.0126  
## area 203.3213 319.6141  
## bedrooms -67317.9617 279091.6700  
## bathrooms 791584.0346 1333607.3037  
## stories 236134.9934 537926.9055  
## mainroad 177052.5819 907119.2757  
## guestroom 95872.0490 693294.6911  
## hotwaterheating 489201.8238 1636648.2875  
## airconditioning 570151.0318 1105278.7632  
## prefarea 463373.7274 1020460.8663  
## furnishingstatus -421826.3838 -111366.7093

# comparing results with removing outliers  
# Since we have many observations , we can remove the outliers from price and area as they seem less significant  
#Removing outliers for price and area columns and converting categorical data to integer-19BDS0144  
df<-House  
boxplot(list(df$price,df$area),horizontal = TRUE,xlab='With outliers' ,col="orange")



df$airconditioning<-as.integer(as.factor(df$airconditioning))  
df$mainroad<-as.integer(as.factor(df$mainroad))  
df$hotwaterheating<-as.integer(as.factor(df$hotwaterheating))  
df$prefarea<-as.integer(as.factor(df$prefarea))  
df$guestroom<-as.integer(as.factor(df$guestroom))  
df$furnishingstatus<- as.integer(as.factor(df$furnishingstatus))  
df$basement<-as.integer(as.factor(df$basement))  
df<- df[-which(df$price %in% out\_price),-which(df$area %in% out\_area)]  
boxplot(list(df$price,df$area),horizontal = TRUE,xlab='Without outliers',col="blue" )



str(df)

## 'data.frame': 530 obs. of 11 variables:  
## $ price : int 9100000 9100000 8960000 8890000 8855000 8750000 8680000 8645000 8645000 8575000 ...  
## $ area : int 6000 6600 8500 4600 6420 4320 7155 8050 4560 8800 ...  
## $ bedrooms : int 4 4 3 3 3 3 3 3 3 3 ...  
## $ bathrooms : int 1 2 2 2 2 1 2 1 2 2 ...  
## $ stories : int 2 2 4 2 2 2 1 1 2 2 ...  
## $ mainroad : int 2 2 2 2 2 2 2 2 2 2 ...  
## $ guestroom : int 1 2 1 2 1 1 2 2 2 1 ...  
## $ hotwaterheating : int 1 1 1 1 1 2 1 1 1 1 ...  
## $ airconditioning : int 1 2 2 2 2 1 2 2 2 2 ...  
## $ prefarea : int 1 2 1 1 2 1 1 1 1 1 ...  
## $ furnishingstatus: int 2 3 1 1 2 2 3 1 1 1 ...

set.seed(1)  
# we will partition into 70% training and 30% validation  
trainrows <- sample(rownames(df), dim(df)[1]\*0.7)  
train <- df[trainrows, ]  
# we will set the difference of the training into the validation set i.e. 30%  
validrows <- setdiff(rownames(df), trainrows)  
valid <- df[validrows, ]

model\_1<-lm(formula= price~area+ bedrooms+bathrooms+stories+mainroad+guestroom+hotwaterheating+airconditioning+prefarea+furnishingstatus ,data= df,subset = trainrows)  
summary(model\_1)

##   
## Call:  
## lm(formula = price ~ area + bedrooms + bathrooms + stories +   
## mainroad + guestroom + hotwaterheating + airconditioning +   
## prefarea + furnishingstatus, data = df, subset = trainrows)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2472049 -589623 -32147 526312 4741527   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.445e+06 5.227e+05 -4.677 4.13e-06 \*\*\*  
## area 2.301e+02 2.544e+01 9.046 < 2e-16 \*\*\*  
## bedrooms 1.536e+05 7.482e+04 2.053 0.04075 \*   
## bathrooms 8.474e+05 1.125e+05 7.534 4.02e-13 \*\*\*  
## stories 4.190e+05 6.506e+04 6.440 3.82e-10 \*\*\*  
## mainroad 4.597e+05 1.517e+05 3.030 0.00262 \*\*   
## guestroom 4.085e+05 1.315e+05 3.107 0.00204 \*\*   
## hotwaterheating 7.312e+05 2.618e+05 2.793 0.00550 \*\*   
## airconditioning 8.245e+05 1.161e+05 7.100 6.69e-12 \*\*\*  
## prefarea 5.811e+05 1.222e+05 4.755 2.87e-06 \*\*\*  
## furnishingstatus -1.711e+05 6.635e+04 -2.578 0.01032 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 945300 on 360 degrees of freedom  
## Multiple R-squared: 0.6514, Adjusted R-squared: 0.6417   
## F-statistic: 67.27 on 10 and 360 DF, p-value: < 2.2e-16

pred<-predict(model\_1,newdata= valid)  
vres<-data.frame(valid$price,pred,residuals=valid$price-pred)  
head(vres) #the difference in valid.price vs pred has decreased

## valid.price pred residuals  
## 20 8855000 6554444 2300556  
## 21 8750000 4549530 4200470  
## 22 8680000 5960917 2719083  
## 23 8645000 5661619 2983381  
## 24 8645000 6124969 2520031  
## 25 8575000 6692005 1882995

#Thus,removing outliers in Price and Area has improved the model accuracy

```

Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.