Unit 3 Homework - Housing regression with AES data

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*.

# load packages

packages <- c("AER", "ggplot2", "PerformanceAnalytics")  
  
sapply(packages, library, character.only = TRUE)

## Loading required package: car

## Loading required package: lmtest

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

## Loading required package: xts

##   
## Attaching package: 'PerformanceAnalytics'

## The following object is masked from 'package:graphics':  
##   
## legend

## $AER  
## [1] "AER" "survival" "sandwich" "lmtest" "zoo"   
## [6] "car" "stats" "graphics" "grDevices" "utils"   
## [11] "datasets" "methods" "base"   
##   
## $ggplot2  
## [1] "ggplot2" "AER" "survival" "sandwich" "lmtest"   
## [6] "zoo" "car" "stats" "graphics" "grDevices"  
## [11] "utils" "datasets" "methods" "base"   
##   
## $PerformanceAnalytics  
## [1] "PerformanceAnalytics" "xts" "ggplot2"   
## [4] "AER" "survival" "sandwich"   
## [7] "lmtest" "zoo" "car"   
## [10] "stats" "graphics" "grDevices"   
## [13] "utils" "datasets" "methods"   
## [16] "base"

# Read in the data

data("HousePrices")  
  
df<- HousePrices

# Quick EDA

summary(df)

## price lotsize bedrooms bathrooms   
## Min. : 25000 Min. : 1650 Min. :1.000 Min. :1.000   
## 1st Qu.: 49125 1st Qu.: 3600 1st Qu.:2.000 1st Qu.:1.000   
## Median : 62000 Median : 4600 Median :3.000 Median :1.000   
## Mean : 68122 Mean : 5150 Mean :2.965 Mean :1.286   
## 3rd Qu.: 82000 3rd Qu.: 6360 3rd Qu.:3.000 3rd Qu.:2.000   
## Max. :190000 Max. :16200 Max. :6.000 Max. :4.000   
## stories driveway recreation fullbase gasheat aircon   
## Min. :1.000 no : 77 no :449 no :355 no :521 no :373   
## 1st Qu.:1.000 yes:469 yes: 97 yes:191 yes: 25 yes:173   
## Median :2.000   
## Mean :1.808   
## 3rd Qu.:2.000   
## Max. :4.000   
## garage prefer   
## Min. :0.0000 no :418   
## 1st Qu.:0.0000 yes:128   
## Median :0.0000   
## Mean :0.6923   
## 3rd Qu.:1.0000   
## Max. :3.0000

str(df)

## 'data.frame': 546 obs. of 12 variables:  
## $ price : num 42000 38500 49500 60500 61000 66000 66000 69000 83800 88500 ...  
## $ lotsize : num 5850 4000 3060 6650 6360 4160 3880 4160 4800 5500 ...  
## $ bedrooms : num 3 2 3 3 2 3 3 3 3 3 ...  
## $ bathrooms : num 1 1 1 1 1 1 2 1 1 2 ...  
## $ stories : num 2 1 1 2 1 1 2 3 1 4 ...  
## $ driveway : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...  
## $ recreation: Factor w/ 2 levels "no","yes": 1 1 1 2 1 2 1 1 2 2 ...  
## $ fullbase : Factor w/ 2 levels "no","yes": 2 1 1 1 1 2 2 1 2 1 ...  
## $ gasheat : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...  
## $ aircon : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 1 1 2 ...  
## $ garage : num 1 0 0 0 0 0 2 0 0 1 ...  
## $ prefer : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...

head(df)

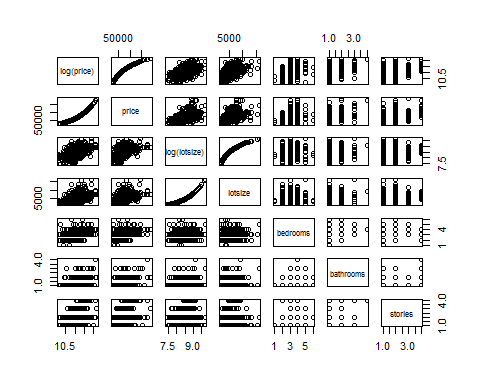
## price lotsize bedrooms bathrooms stories driveway recreation fullbase  
## 1 42000 5850 3 1 2 yes no yes  
## 2 38500 4000 2 1 1 yes no no  
## 3 49500 3060 3 1 1 yes no no  
## 4 60500 6650 3 1 2 yes yes no  
## 5 61000 6360 2 1 1 yes no no  
## 6 66000 4160 3 1 1 yes yes yes  
## gasheat aircon garage prefer  
## 1 no no 1 no  
## 2 no no 0 no  
## 3 no no 0 no  
## 4 no no 0 no  
## 5 no no 0 no  
## 6 no yes 0 no

### Looks like we have a few factors that are 2 levels and not boolean. Although R will fix this, lets fix that:

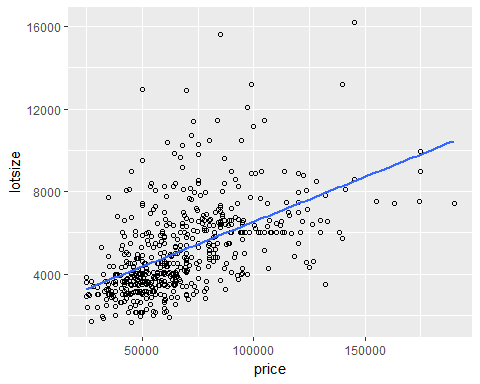
df$driveway <- as.numeric(df$driveway =='yes')  
df$recreation<- as.numeric(df$recreation =='yes')  
df$fullbase <- as.numeric(df$fullbase =='yes')  
df$gasheat <- as.numeric(df$gasheat =='yes')  
df$aircon <- as.numeric(df$aircon == 'yes')  
df$prefer <- as.numeric(df$prefer == 'yes')

### let's look at some breakdowns of the data:

df\_names <- names(df)  
  
#scatterplotMatrix(~price + lotsize + bedrooms + bathrooms , data = df)  
pairs(~log(price) +price + log(lotsize) + lotsize + bedrooms + bathrooms +stories , data = df)

 ### the scatterplot matrix is interesting,let's take a deeper look:

par(mfrow = c(3,1))  
  
# plot price versus lot size  
ggplot(df, aes(x = price, y = lotsize)) + geom\_point(shape = 1) + geom\_smooth(method = lm, se = FALSE)



# plots log(price) versus lot size  
  
ggplot(df, aes(x = log(price), y = lotsize)) + geom\_point(shape = 1) + geom\_smooth(method = lm, se = FALSE)

