Sample R code

Here's some sample R code that takes the Brooklyn housing data in the preceding exercise, and cleans and explores it a bit. (The exercise asks you to do this for Manhattan.)

```
# Author: Benjamin Reddy

library(plyr)

require(gdata)
bk <- read.xls("rollingsales_brooklyn.xls",pattern="BOROUGH")
head(bk)
summary(bk)

bk$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","",

                                bk$SALE.PRICE))
count(is.na(bk$SALE.PRICE.N))

names(bk) <- tolower(names(bk))

## clean/format the data with regular expressions
bk$gross.sqft <- as.numeric(gsub("[^[:digit:]]","",
                            bk$gross.square.feet))
bk$land.sqft <- as.numeric(gsub("[^[:digit:]]","",
                            bk$land.square.feet))

bk$sale.date <- as.Date(bk$sale.date)
bk$year.built <- as.numeric(as.character(bk$year.built))

## do a bit of exploration to make sure there's not anything
## weird going on with sale prices
attach(bk)

hist(sale.price.n)
hist(sale.price.n[sale.price.n>0])
hist(gross.sqft[sale.price.n==0])

detach(bk)

## keep only the actual sales
bk.sale <- bk[bk$sale.price.n!=0,]

plot(bk.sale$gross.sqft,bk.sale$sale.price.n)
plot(log(bk.sale$gross.sqft),log(bk.sale$sale.price.n))

## for now, let's look at 1-, 2-, and 3-family homes
bk.homes <- bk.sale[which(grepl("FAMILY",
            bk.sale$building.class.category)),]
plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))

bk.homes[which(bk.homes$sale.price.n<100000),]
    [order(bk.homes[which(bk.homes$sale.price.n<100000),]
           $sale.price.n),]

## remove outliers that seem like they weren't actual sales
bk.homes$outliers <- (log(bk.homes$sale.price.n) <=5) + 0
bk.homes <- bk.homes[which(bk.homes$outliers==0),]

plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
```