EDA of Brooklyn Housing Data

Jon Kinsey

Thu Dec 4 00:09:52 2014

```
# Brooklyn housing data - rolling sales (Doing Data Science)
#
library(ggplot2)
library(plyr)
require(gdata)
```

```
## Loading required package: gdata
## gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
##
## gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
##
## Attaching package: 'gdata'
##
## The following object is masked from 'package:stats':
##
##     nobs
##
## The following object is masked from 'package:utils':
##
##     object.size
```

```
bk <- read.xls("/Users/Jon/Desktop/R-Projects/Brooklyn/rollingsales_brooklyn.xls",pattern="BOROUGH")
names(bk)
```

```
##  [1] "BOROUGH"                       "NEIGHBORHOOD"
##  [3] "BUILDING.CLASS.CATEGORY"       "TAX.CLASS.AT.PRESENT"
##  [5] "BLOCK"                         "LOT"
##  [7] "EASE.MENT"                     "BUILDING.CLASS.AT.PRESENT"
##  [9] "ADDRESS"                       "APART.MENT.NUMBER"
## [11] "ZIP.CODE"                      "RESIDENTIAL.UNITS"
## [13] "COMMERCIAL.UNITS"              "TOTAL.UNITS"
## [15] "LAND.SQUARE.FEET"              "GROSS.SQUARE.FEET"
## [17] "YEAR.BUILT"                    "TAX.CLASS.AT.TIME.OF.SALE"
## [19] "BUILDING.CLASS.AT.TIME.OF.SALE" "SALE.PRICE"
## [21] "SALE.DATE"
```

```
head(bk)
```

```
##   BOROUGH              NEIGHBORHOOD
## 1       3
## 2       3
## 3       3
## 4       3
## 5       3
## 6       3
##                       BUILDING.CLASS.CATEGORY TAX.CLASS.AT.PRESENT BLOCK
```

```
## 1 15  CONDOS - 2-10 UNIT RESIDENTIAL                                    814
## 2 15  CONDOS - 2-10 UNIT RESIDENTIAL                                    814
## 3 15  CONDOS - 2-10 UNIT RESIDENTIAL                                   1967
## 4 15  CONDOS - 2-10 UNIT RESIDENTIAL                                   1967
## 5 15  CONDOS - 2-10 UNIT RESIDENTIAL                                   1967
## 6 15  CONDOS - 2-10 UNIT RESIDENTIAL                                   1967
##    LOT EASE.MENT BUILDING.CLASS.AT.PRESENT
## 1 1103        NA
## 2 1105        NA
## 3 1401        NA
## 4 1402        NA
## 5 1403        NA
## 6 1404        NA
##                                   ADDRESS APART.MENT.NUMBER ZIP.CODE
## 1 342 53RD    STREET                                          11220
## 2 342 53RD    STREET                                          11220
## 3 290 GREENE AVE                                              11238
## 4 290 GREENE AVE                                              11238
## 5 290 GREENE AVE                                              11238
## 6 290 GREENE AVE                                              11238
##   RESIDENTIAL.UNITS COMMERCIAL.UNITS TOTAL.UNITS LAND.SQUARE.FEET
## 1                 0                0           0                0
## 2                 0                0           0                0
## 3                 0                0           0                0
## 4                 0                0           0                0
## 5                 0                0           0                0
## 6                 0                0           0                0
##   GROSS.SQUARE.FEET YEAR.BUILT TAX.CLASS.AT.TIME.OF.SALE
## 1                 0          0                        2
## 2                 0          0                        2
## 3                 0          0                        2
## 4                 0          0                        2
## 5                 0          0                        2
## 6                 0          0                        2
##   BUILDING.CLASS.AT.TIME.OF.SALE SALE.PRICE  SALE.DATE
## 1                             R1   $403,572 2013-07-09
## 2                             R1   $218,010 2013-07-12
## 3                             R1   $952,311 2013-04-25
## 4                             R1   $842,692 2013-04-25
## 5                             R1   $815,288 2013-04-25
## 6                             R1   $815,288 2013-04-25
```

```
summary(bk)
```

```
##     BOROUGH                     NEIGHBORHOOD
##  Min.   :3   BEDFORD STUYVESANT       : 1699
##  1st Qu.:3   EAST NEW YORK            : 1394
##  Median :3   BOROUGH PARK             : 1020
##  Mean   :3   BUSHWICK                 :  898
##  3rd Qu.:3   CROWN HEIGHTS            :  886
##  Max.   :3   PARK SLOPE               :  848
##              (Other)                  :16628
##                                BUILDING.CLASS.CATEGORY
##  02  TWO FAMILY HOMES                        :5776
```

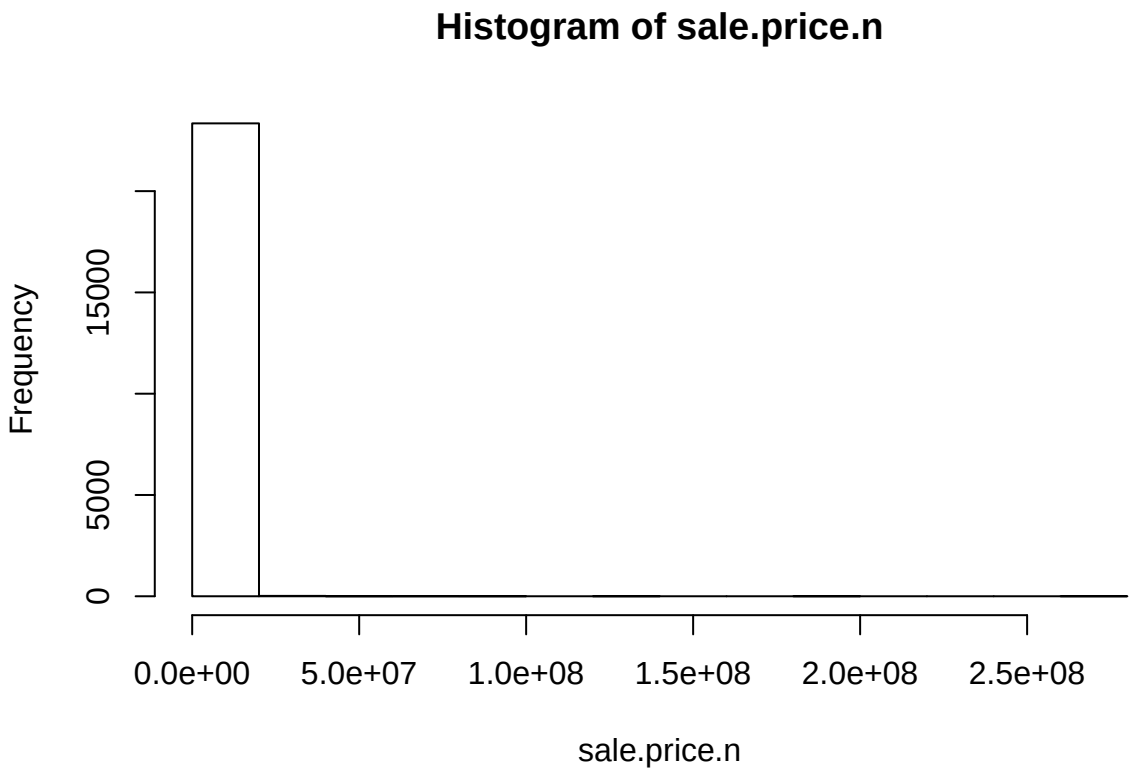
```
##  01  ONE FAMILY HOMES                          :2890
##  13  CONDOS - ELEVATOR APARTMENTS             :2739
##  03  THREE FAMILY HOMES                       :2255
##  10  COOPS - ELEVATOR APARTMENTS              :2129
##  07  RENTALS - WALKUP APARTMENTS              :1755
##  (Other)                                      :5829
##  TAX.CLASS.AT.PRESENT     BLOCK           LOT        EASE.MENT
##  1      :10976        Min.   :  20   Min.   :   1   Mode:logical
##  2      : 6070        1st Qu.:1638   1st Qu.:  22   NA's:23373
##  4      : 2445        Median :3839   Median :  48
##  2A     : 1512        Mean   :3984   Mean   : 305
##  2C     : 1024        3rd Qu.:6259   3rd Qu.: 142
##  1B     :  422        Max.   :8955   Max.   :9039
##  (Other):  924
##  BUILDING.CLASS.AT.PRESENT
##  R4     : 2703
##  C0     : 2258
##  D4     : 2125
##  B1     : 2080
##  B3     : 1229
##  B2     : 1115
##  (Other):11863
##                                        ADDRESS          APART.MENT.NUMBER
##  163 WASHINGTON AVENUE                    :  106               :17632
##  205 WATER STREET                         :   76   4           :  204
##  380 COZINE AVENUE                        :   65   6           :  183
##  34 NORTH 7TH   STREET                    :   63   3           :  155
##  12399 FLATLANDS AVENUE                   :   62   2           :  144
##  306 GOLD STREET                          :   62   1           :  125
##  (Other)                                  :22939   (Other)     : 4930
##     ZIP.CODE     RESIDENTIAL.UNITS COMMERCIAL.UNITS  TOTAL.UNITS
##  Min.   :    0   Min.   :  0.0     Min.   :  0.0    Min.   :  0.0
##  1st Qu.:11209   1st Qu.:  1.0     1st Qu.:  0.0    1st Qu.:  1.0
##  Median :11218   Median :  1.0     Median :  0.0    Median :  1.0
##  Mean   :11211   Mean   :  2.2     Mean   :  0.2    Mean   :  2.4
##  3rd Qu.:11230   3rd Qu.:  2.0     3rd Qu.:  0.0    3rd Qu.:  2.0
##  Max.   :11416   Max.   :509.0     Max.   :222.0    Max.   :509.0
##
##  LAND.SQUARE.FEET GROSS.SQUARE.FEET   YEAR.BUILT
##  0      : 8027    0      : 8934     Min.   :   0
##  2,000  : 2201    3,000  :  230     1st Qu.:1901
##  2,500  : 1149    3,600  :  189     Median :1925
##  1,800  :  597    2,400  :  185     Mean   :1681
##  4,000  :  474    2,700  :  146     3rd Qu.:1950
##  3,000  :  307    3,300  :  139     Max.   :2013
##  (Other):10618    (Other):13550
##  TAX.CLASS.AT.TIME.OF.SALE BUILDING.CLASS.AT.TIME.OF.SALE    SALE.PRICE
##  Min.   :1.00              R4     : 2739                  $0      : 8791
##  1st Qu.:1.00              C0     : 2255                  $10     :  241
##  Median :1.00              D4     : 2125                  $700,000:  138
##  Mean   :1.71              B1     : 2070                  $650,000:  129
##  3rd Qu.:2.00              B3     : 1230                  $300,000:  120
##  Max.   :4.00              B2     : 1115                  $600,000:  120
##                            (Other):11839                  (Other) :13834
```

```
##       SALE.DATE
##  2012-09-27:  675
##  2012-12-27:  245
##  2012-12-20:  222
##  2013-03-22:  204
##  2012-12-31:  179
##  2012-12-19:  178
##  (Other)   :21670
```

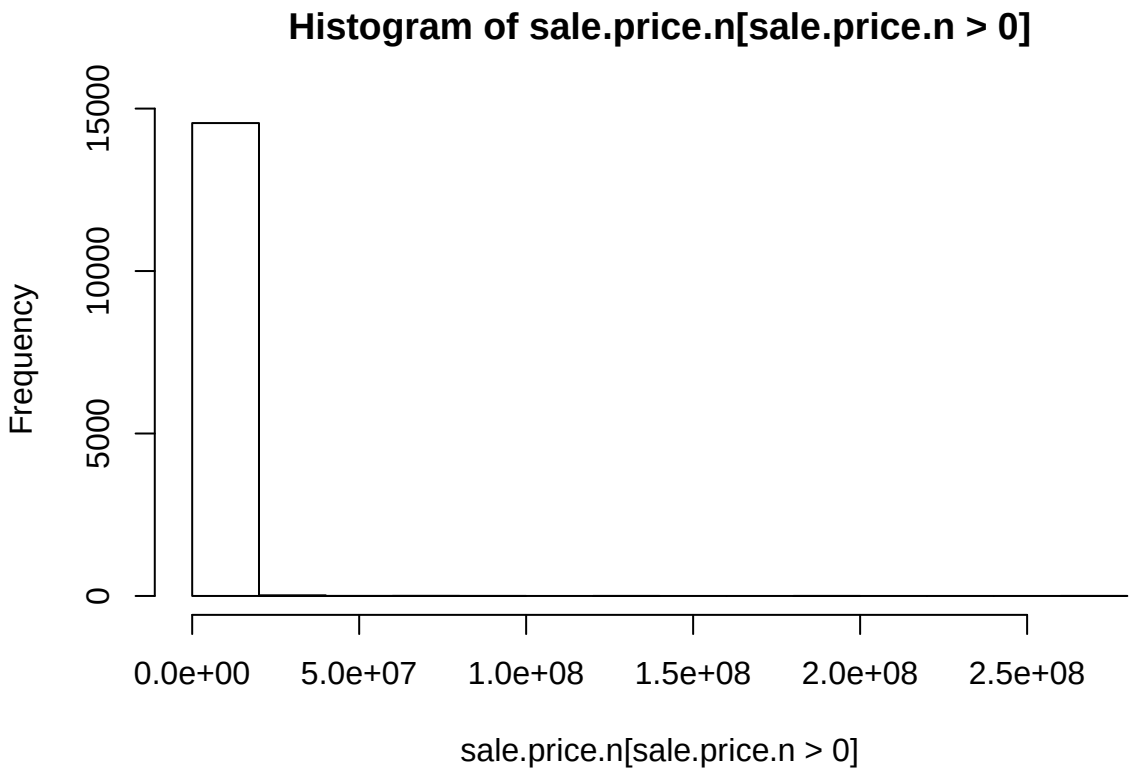
```
bk$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","",bk$SALE.PRICE))
# count using plyr package,
# see http://www.inside-r.org/packages/cran/plyr/docs/count
count(is.na(bk$SALE.PRICE.N))
```

```
##       x  freq
## 1 FALSE 23373
```

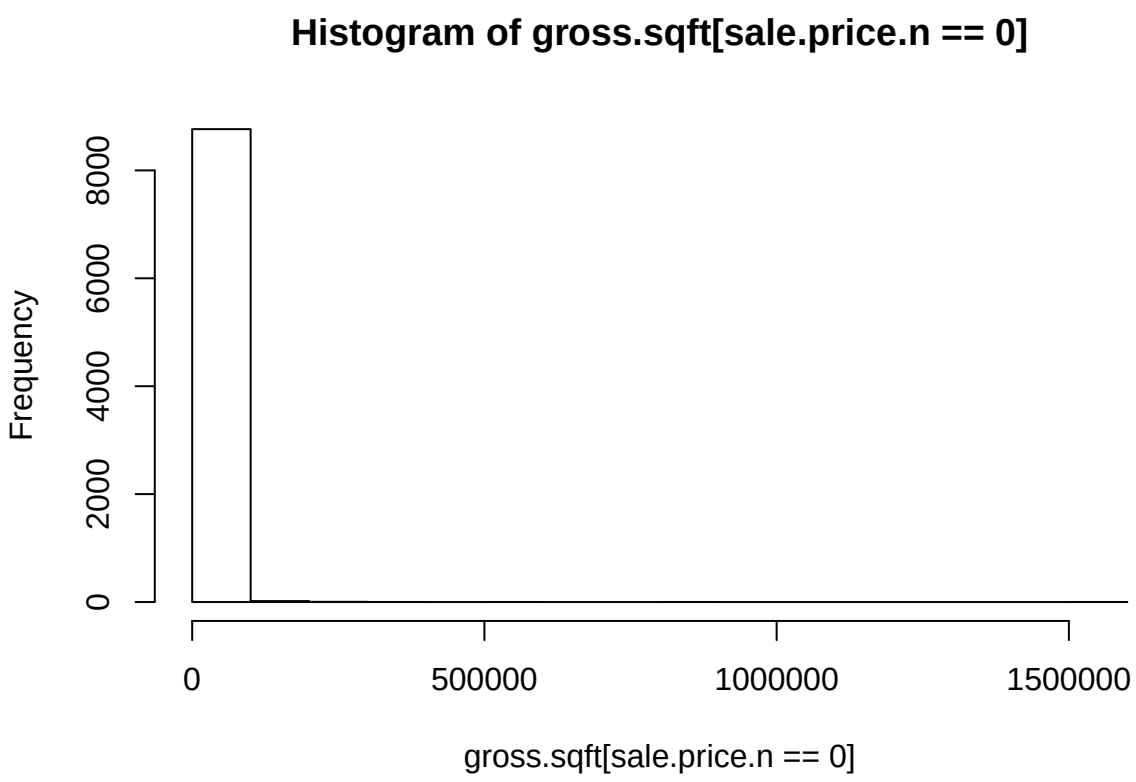
```
# convert to lower case using chartr package,
# see http://www.inside-r.org/r-doc/base/tolower
names(bk) <- tolower(names(bk))

## 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 EDA to make sure sale prices look reasonable
attach(bk)
hist(sale.price.n)
```



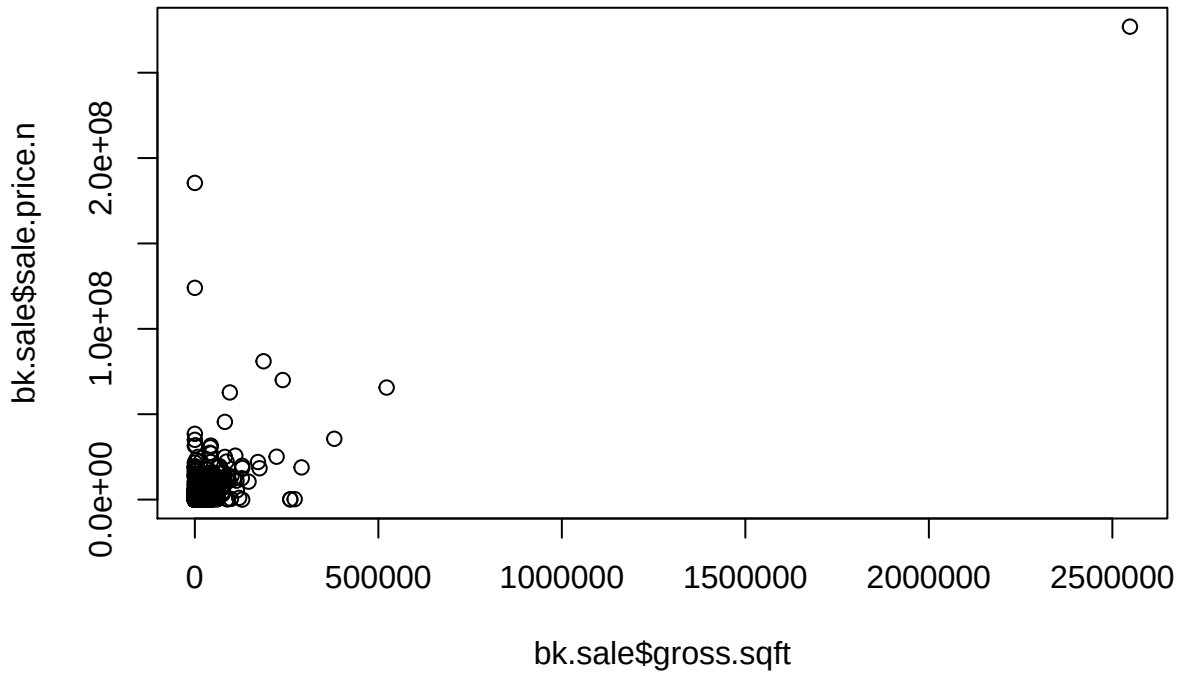

```
hist(sale.price.n[sale.price.n>0])
```



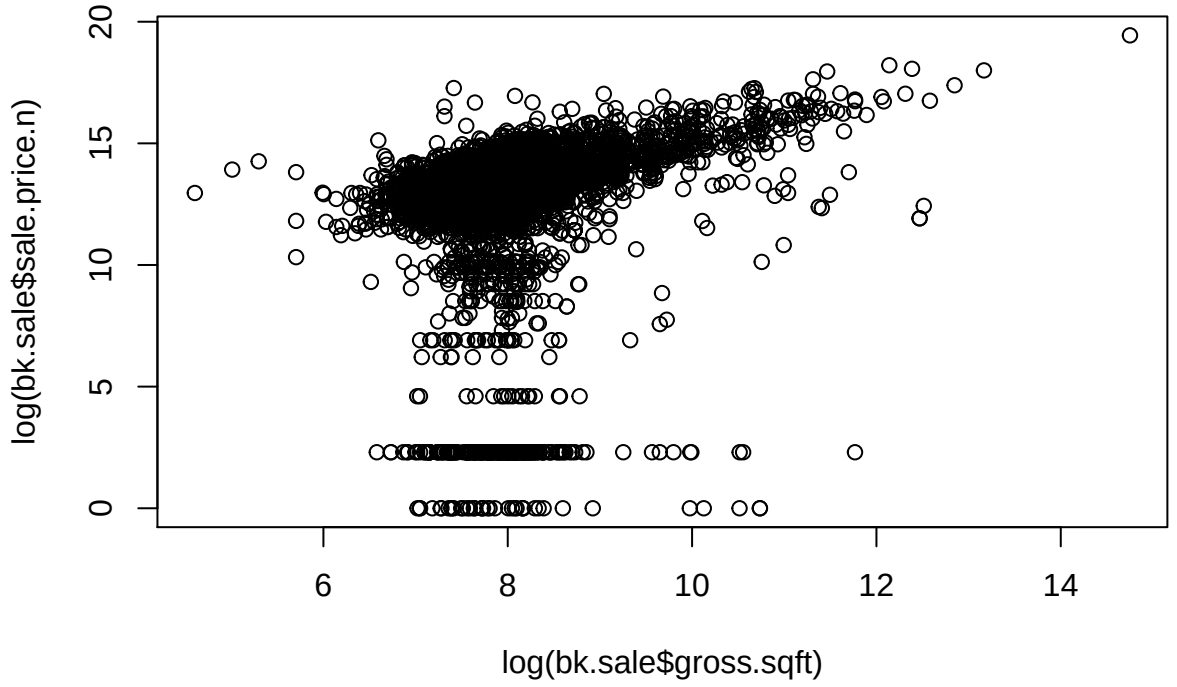
```
hist(gross.sqft[sale.price.n==0])
```



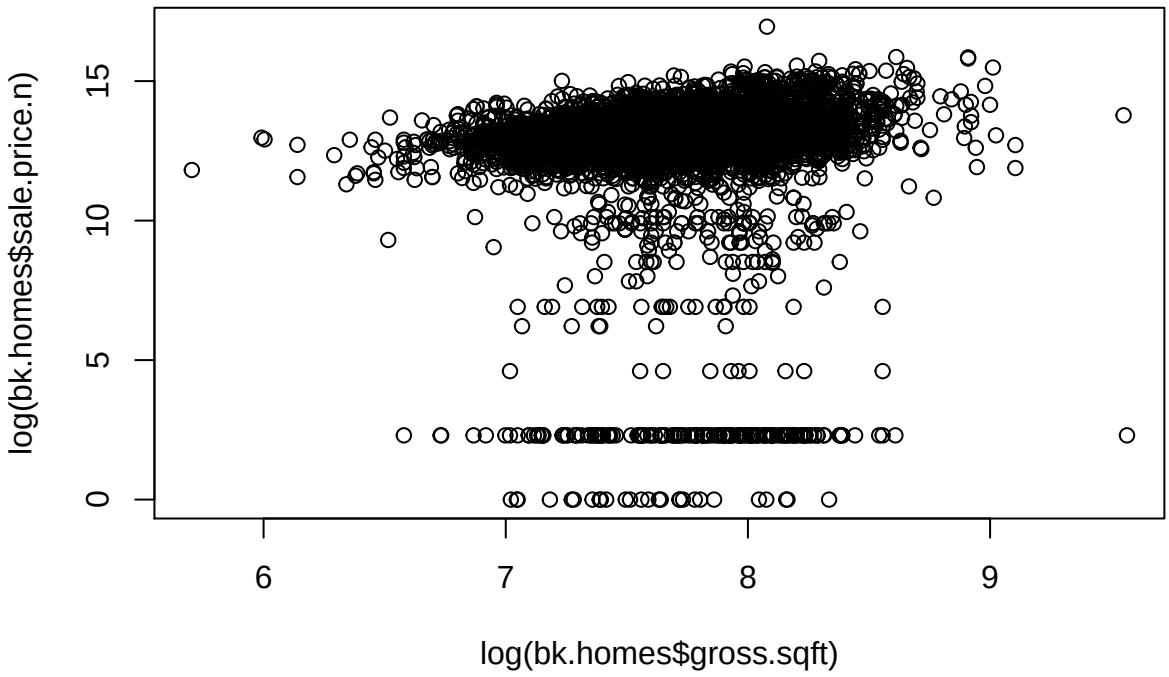

```
detach(bk)

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

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



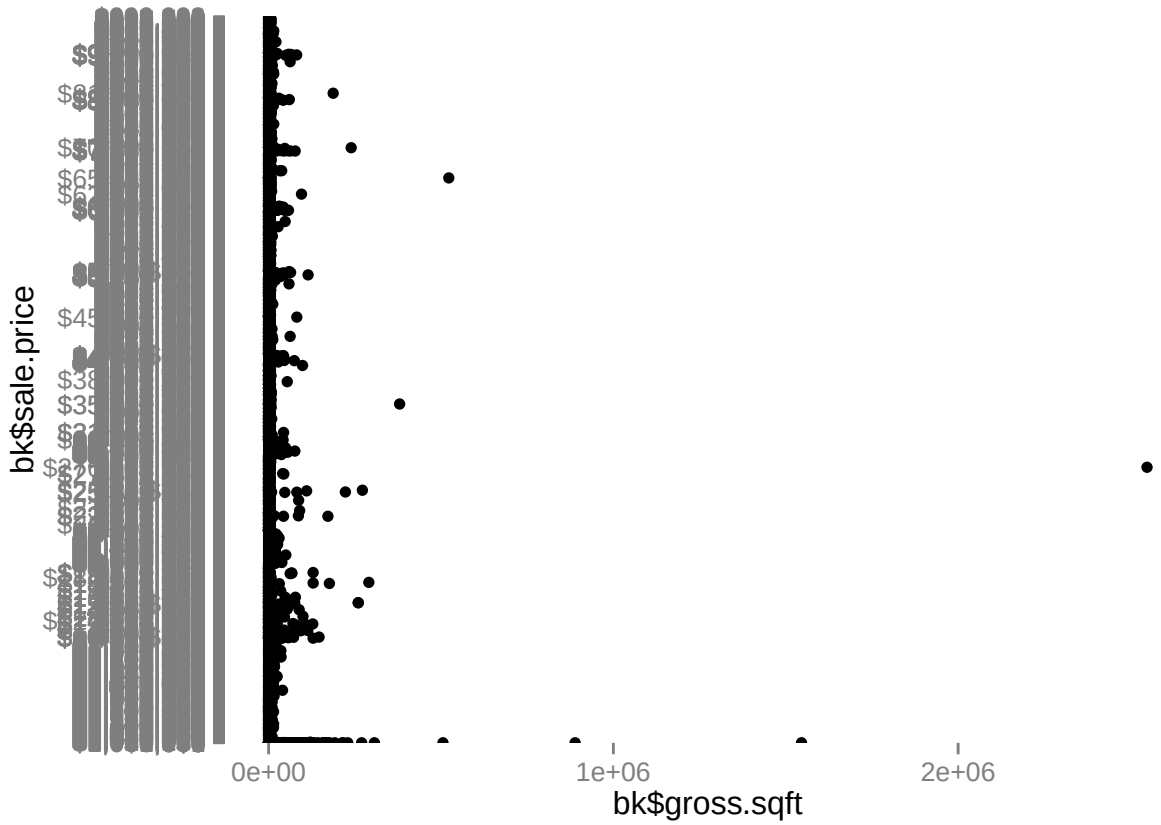

```
plot(log(bk.sale$gross.sqft),log(bk.sale$sale.price.n))
```



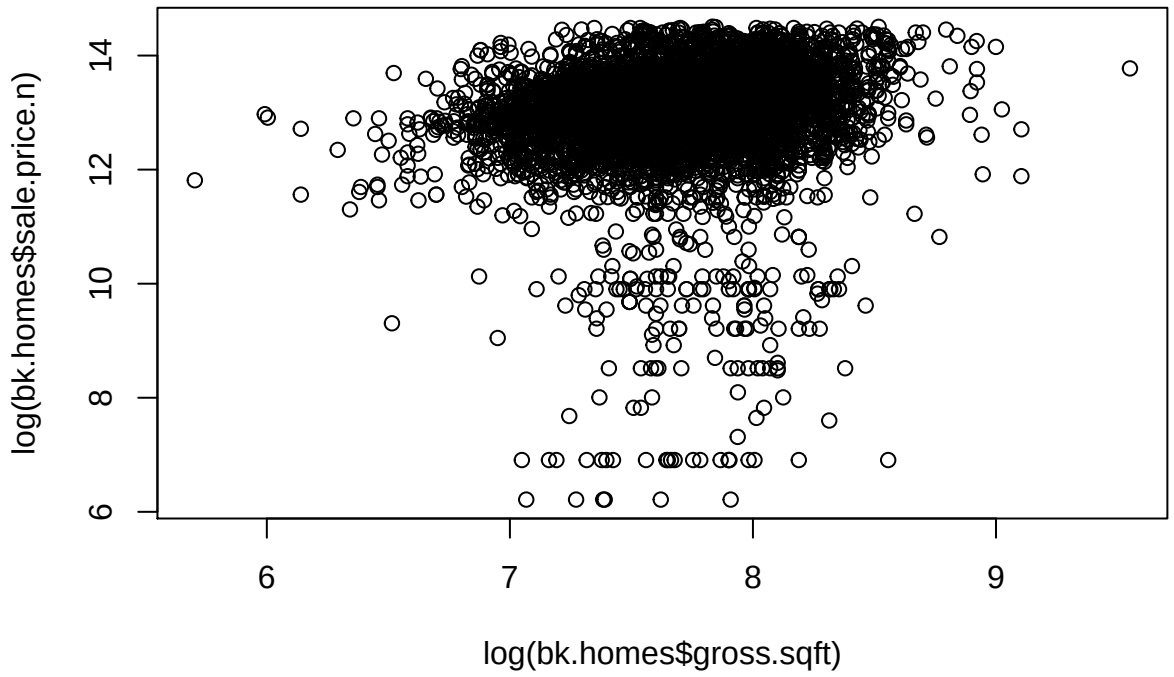

```
# look at 1,2, and 3-family homes
bk.homes <- bk.sale[which(grepl("FAMILY",bk.sale$building.class.category)),]
plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
```



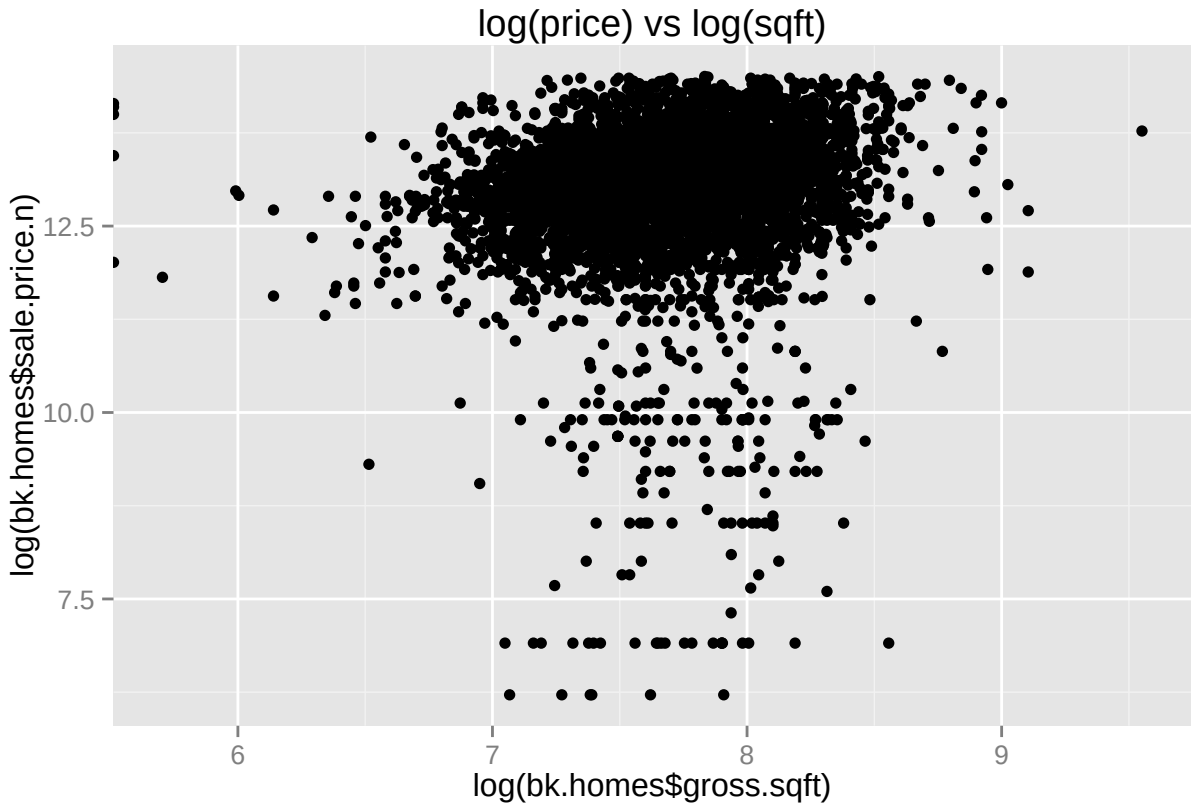

```
# To sort a data frame in R, use the order( ) function.
# By default, sorting is ASCENDING, http://www.statmethods.net/management/sorting.html
bk.homes <- bk.homes[which(bk.homes$sale.price.n<2000000),]
bk.homes <- bk.homes[order(bk.homes[which(bk.homes$sale.price.n<2000000),]$sale.price.n),]

qplot(bk$gross.sqft, bk$sale.price)
```




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




```
qplot(log(bk.homes$gross.sqft), log(bk.homes$sale.price.n)) +
  ggtitle('log(price) vs log(sqft)')
```




```
# zoom in on largest grouping
qplot(log(bk.homes$gross.sqft), log(bk.homes$sale.price.n)) +
  scale_x_continuous(limits = c(6.5,8.5)) +
```

```
  scale_y_continuous(limits = c(11,15)) +
  ggtitle('log(price) vs log(sqft)')
```

```
## Warning: Removed 268 rows containing missing values (geom_point).
```