Exploratory Data Analysis

<https://github.com/oreillymedia/doing_data_science> 에서 전체 데이터 다운

# 1. 뉴욕타임즈 광고노출 및 클릭 데이터

nyt1.csv,..., nyt31.csv는 2012년 5월 뉴욕타임즈 홈페이지에서 일별 광고노출횟수와 클릭수를 담은 데이터로 각 행은 사용자를 의미하고 나이, 성별(0:여성, 1:남성), 노출횟수(impression), 클릭횟수, 로그인 횟수의 변수들로 구성됨

## 데이터 탐색 및 정리

# Author: Maura Fitzgerald   
data1 <- read.csv(url("http://stat.columbia.edu/~rachel/datasets/nyt1.csv"))  
  
# 연령대 범주화   
head(data1)

## Age Gender Impressions Clicks Signed\_In  
## 1 36 0 3 0 1  
## 2 73 1 3 0 1  
## 3 30 0 3 0 1  
## 4 49 1 3 0 1  
## 5 47 1 11 0 1  
## 6 47 0 11 1 1

data1$agecat <-cut(data1$Age,c(-Inf,0,18,24,34,44,54,64,Inf))  
  
# 데이터 요약 정보  
summary(data1)

## Age Gender Impressions Clicks   
## Min. : 0.00 Min. :0.000 Min. : 0.000 Min. :0.00000   
## 1st Qu.: 0.00 1st Qu.:0.000 1st Qu.: 3.000 1st Qu.:0.00000   
## Median : 31.00 Median :0.000 Median : 5.000 Median :0.00000   
## Mean : 29.48 Mean :0.367 Mean : 5.007 Mean :0.09259   
## 3rd Qu.: 48.00 3rd Qu.:1.000 3rd Qu.: 6.000 3rd Qu.:0.00000   
## Max. :108.00 Max. :1.000 Max. :20.000 Max. :4.00000   
##   
## Signed\_In agecat   
## Min. :0.0000 (-Inf,0]:137106   
## 1st Qu.:0.0000 (34,44] : 70860   
## Median :1.0000 (44,54] : 64288   
## Mean :0.7009 (24,34] : 58174   
## 3rd Qu.:1.0000 (54,64] : 44738   
## Max. :1.0000 (18,24] : 35270   
## (Other) : 48005

# 연령대별로 통계량 보기  
library("doBy")   
siterange <- function(x){c(length(x), min(x), mean(x), max(x))}   
summaryBy(Age~agecat, data =data1, FUN=siterange)

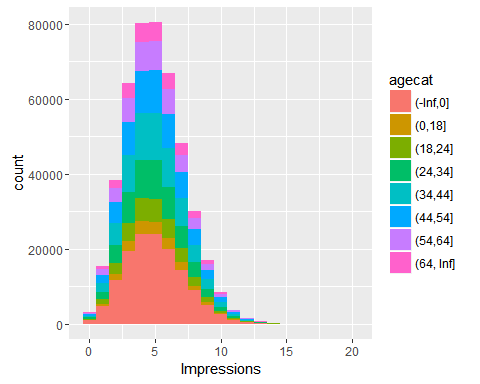
## agecat Age.FUN1 Age.FUN2 Age.FUN3 Age.FUN4  
## 1 (-Inf,0] 137106 0 0.00000 0  
## 2 (0,18] 19252 7 16.03350 18  
## 3 (18,24] 35270 19 21.26904 24  
## 4 (24,34] 58174 25 29.50335 34  
## 5 (34,44] 70860 35 39.49468 44  
## 6 (44,54] 64288 45 49.49258 54  
## 7 (54,64] 44738 55 59.49819 64  
## 8 (64, Inf] 28753 65 72.98870 108

# 로그인한 사용자에게만 나이와 성별 부여  
summaryBy(Gender+Signed\_In+Impressions+Clicks~agecat, data =data1)

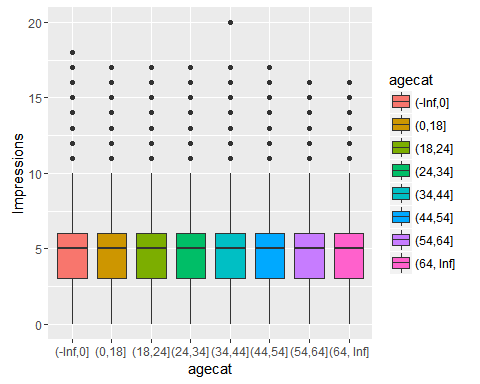
## agecat Gender.mean Signed\_In.mean Impressions.mean Clicks.mean  
## 1 (-Inf,0] 0.0000000 0 4.999657 0.14207985  
## 2 (0,18] 0.6421151 1 4.998961 0.13105132  
## 3 (18,24] 0.5338531 1 5.006635 0.04845478  
## 4 (24,34] 0.5321621 1 4.993829 0.05048647  
## 5 (34,44] 0.5316963 1 5.021507 0.05167937  
## 6 (44,54] 0.5289790 1 5.010406 0.05027377  
## 7 (54,64] 0.5361885 1 5.022308 0.10183736  
## 8 (64, Inf] 0.3632664 1 5.012347 0.15128856

연령이 음수인 데이터 확인: 성별 x, 로그인 x

# 연령대별 광고노출횟수  
library(ggplot2)   
ggplot(data1, aes(x=Impressions, fill=agecat))+  
 geom\_histogram(binwidth=1)



ggplot(data1, aes(x=agecat, y=Impressions, fill=agecat))+  
 geom\_boxplot()

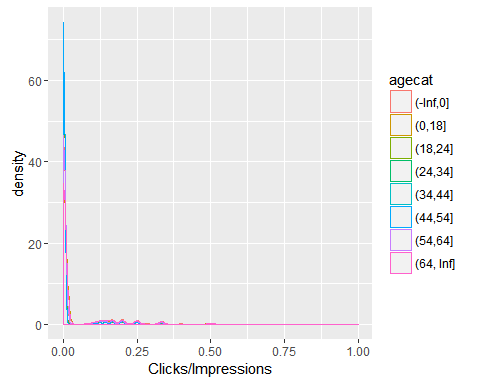


클릭률 생성: 광고노출이 없는 경우 클릭 무시

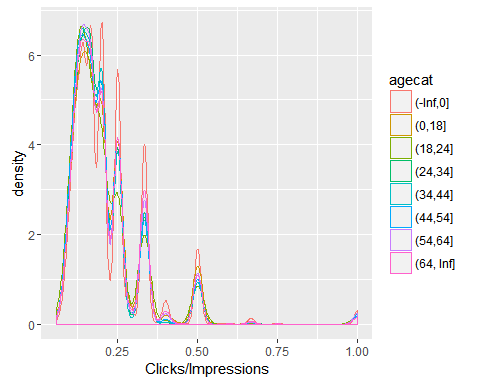
# 광고노출횟수를 음수와 양수로 범주화  
data1$hasimps <-cut(data1$Impressions,c(-Inf,0,Inf))   
summaryBy(Clicks~hasimps, data =data1, FUN=siterange)

## hasimps Clicks.FUN1 Clicks.FUN2 Clicks.FUN3 Clicks.FUN4  
## 1 (-Inf,0] 3066 0 0.00000000 0  
## 2 (0, Inf] 455375 0 0.09321768 4

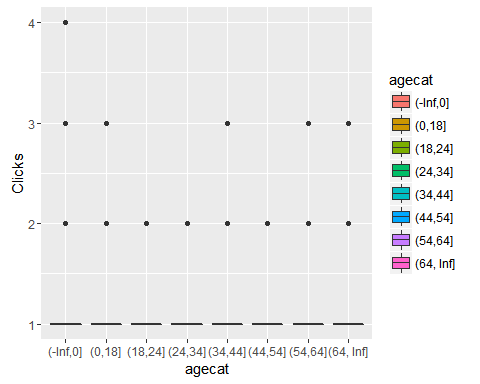
# 광고노출횟수>0인 경우 클릭횟수/광고노출횟수  
ggplot(subset(data1, Impressions>0), aes(x=Clicks/Impressions, colour=agecat)) +   
 geom\_density()



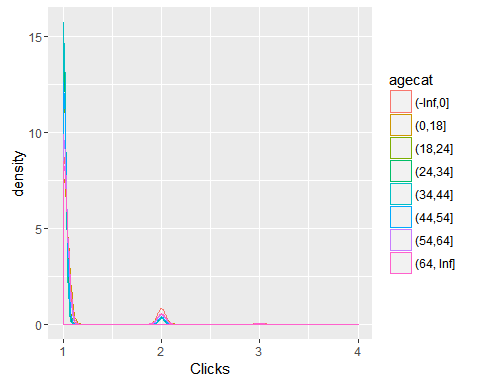
ggplot(subset(data1, Clicks>0), aes(x=Clicks/Impressions, colour=agecat)) +   
 geom\_density()



# 광고노출횟수>0인 경우 클릭수  
ggplot(subset(data1, Clicks>0), aes(x=agecat, y=Clicks, fill=agecat)) +   
 geom\_boxplot()



ggplot(subset(data1, Clicks>0), aes(x=Clicks, colour=agecat)) +   
 geom\_density()



# 광고노출 유무, 클릭 유무 범주화  
data1$scode[data1$Impressions==0] <- "NoImps"   
data1$scode[data1$Impressions >0] <- "Imps"   
data1$scode[data1$Clicks >0] <- "Clicks"  
  
# 범주형 변수로 변환  
data1$scode <- factor(data1$scode)   
head(data1)

## Age Gender Impressions Clicks Signed\_In agecat hasimps scode  
## 1 36 0 3 0 1 (34,44] (0, Inf] Imps  
## 2 73 1 3 0 1 (64, Inf] (0, Inf] Imps  
## 3 30 0 3 0 1 (24,34] (0, Inf] Imps  
## 4 49 1 3 0 1 (44,54] (0, Inf] Imps  
## 5 47 1 11 0 1 (44,54] (0, Inf] Imps  
## 6 47 0 11 1 1 (44,54] (0, Inf] Clicks

# 테이블 생성  
clen <- function(x){c(length(x))}   
etable<-summaryBy(Impressions~scode+Gender+agecat, data = data1, FUN=clen)

## 연습문제

nyt2,..., nyt31에 대하여 유사한 분석을 하고 여러가지 지표와 분포를 시간에 따라 시각화 하시오. 단 분산처리를 염두에 두어 각 날짜별로 데이터를 불러와 처리한 후 가공된 데이터를 데이터 프레임으로 저장하시오.

# 2. RealDirect 데이터

## 데이터 탐색 및 정리

2012년 8월부터 2013년 8월 부루클린 데이터에 대한 기초적인 분석

# Author: Benjamin Reddy  
require(gdata)

## Loading required package: gdata

## gdata: Unable to locate valid perl interpreter  
## gdata:   
## gdata: read.xls() will be unable to read Excel XLS and XLSX files  
## gdata: unless the 'perl=' argument is used to specify the location  
## gdata: of a valid perl intrpreter.  
## gdata:   
## gdata: (To avoid display of this message in the future, please  
## gdata: ensure perl is installed and available on the executable  
## gdata: search path.)

## gdata: Unable to load perl libaries needed by read.xls()  
## gdata: to support 'XLX' (Excel 97-2004) files.

##

## gdata: Unable to load perl libaries needed by read.xls()  
## gdata: to support 'XLSX' (Excel 2007+) files.

##

## gdata: Run the function 'installXLSXsupport()'  
## gdata: to automatically download and install the perl  
## gdata: libaries needed to support Excel XLS and XLSX formats.

##   
## Attaching package: 'gdata'

## The following object is masked from 'package:stats':  
##   
## nobs

## The following object is masked from 'package:utils':  
##   
## object.size

setwd("D:/doingdatascience/dds\_datasets")  
bk <- read.xls("rollingsales\_brooklyn.xls",   
 pattern="BOROUGH", perl = "C:/Perl64/bin/perl.exe")   
  
# 변수 살펴보기  
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

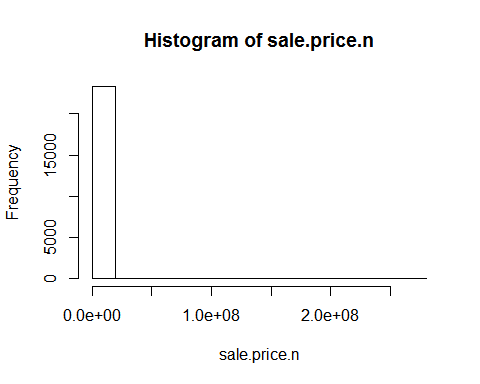
summary(bk)

## 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. : 1.0 Mode:logical   
## 2 : 6070 1st Qu.:1638 1st Qu.: 22.0 NA's:23373   
## 4 : 2445 Median :3839 Median : 48.0   
## 2A : 1512 Mean :3984 Mean : 305.4   
## 2C : 1024 3rd Qu.:6259 3rd Qu.: 142.0   
## 1B : 422 Max. :8955 Max. :9039.0   
## (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. : 0.000 Min. : 0.0000 Min. : 0.00   
## 1st Qu.:11209 1st Qu.: 1.000 1st Qu.: 0.0000 1st Qu.: 1.00   
## Median :11218 Median : 1.000 Median : 0.0000 Median : 1.00   
## Mean :11211 Mean : 2.156 Mean : 0.1973 Mean : 2.37   
## 3rd Qu.:11230 3rd Qu.: 2.000 3rd Qu.: 0.0000 3rd Qu.: 2.00   
## Max. :11416 Max. :509.000 Max. :222.0000 Max. :509.00   
##   
## 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. :1.000 R4 : 2739 $0 : 8791   
## 1st Qu.:1.000 C0 : 2255 $10 : 241   
## Median :1.000 D4 : 2125 $700,000: 138   
## Mean :1.705 B1 : 2070 $650,000: 129   
## 3rd Qu.:2.000 B3 : 1230 $300,000: 120   
## Max. :4.000 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

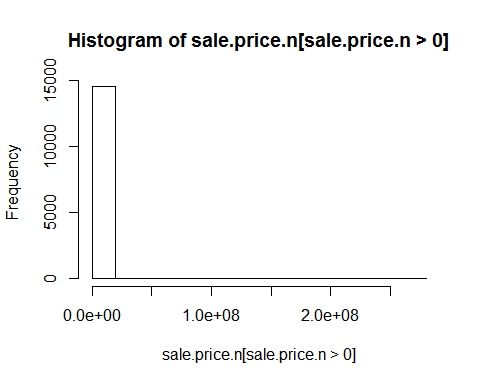
# 데이터 정리  
bk$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","", bk$SALE.PRICE))   
sum(is.na(bk$SALE.PRICE.N))

## [1] 0

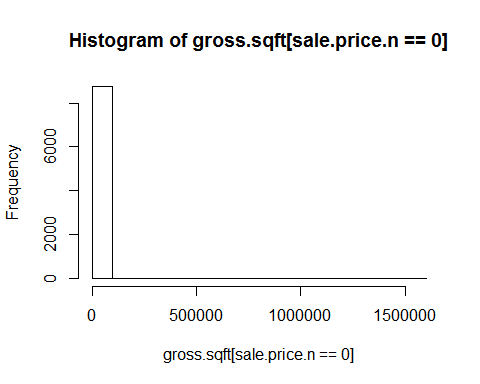
names(bk) <- tolower(names(bk))   
  
bk$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", bk$gross.square.feet))   
bk$land.sqft <- as.numeric(gsub("[^[:digit:]]","", bk$land.square.feet))  
bk$sale.date <- as.Date(bk$sale.date)   
bk$year.built <- as.numeric(as.character(bk$year.built))  
  
# 판매가에 이상이 없는지 탐색  
attach(bk)  
hist(sale.price.n)



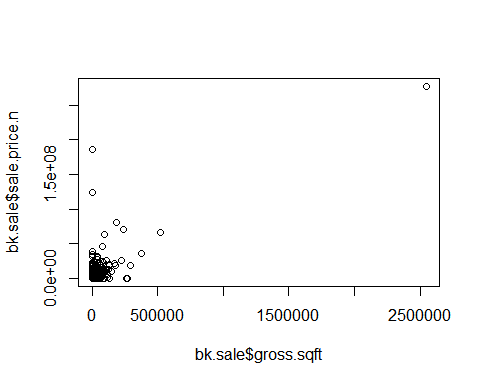
hist(sale.price.n[sale.price.n>0])



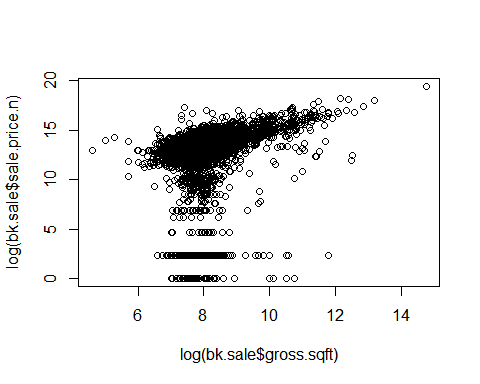
hist(gross.sqft[sale.price.n==0])



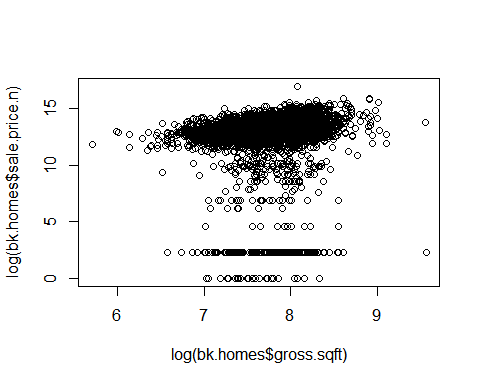
detach(bk)  
  
# 실제 판매가 이루어진것만   
bk.sale <- bk[bk$sale.price.n!=0,]  
plot(bk.sale$gross.sqft, bk.sale$sale.price.n)



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



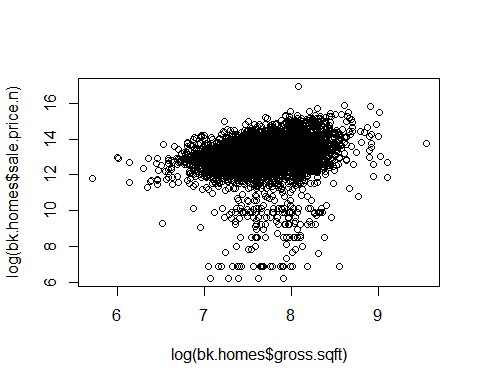
# 단독, 2가구, 3가구 주택  
bk.homes <- bk.sale[which(grepl("FAMILY", bk.sale$building.class.category)),]  
plot(log(bk.homes$gross.sqft), log(bk.homes$sale.price.n))



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

## borough neighborhood  
## 1201 3 BEDFORD STUYVESANT   
## 1451 3 BEDFORD STUYVESANT   
## 2870 3 BENSONHURST   
## 2937 3 BENSONHURST   
## 3155 3 BENSONHURST   
## 5631 3 BROWNSVILLE   
## building.class.category tax.class.at.present  
## 1201 01 ONE FAMILY HOMES 1  
## 1451 02 TWO FAMILY HOMES 1  
## 2870 01 ONE FAMILY HOMES 1  
## 2937 02 TWO FAMILY HOMES 1  
## 3155 03 THREE FAMILY HOMES 1  
## 5631 01 ONE FAMILY HOMES 1  
## block lot ease.ment building.class.at.present  
## 1201 1706 23 NA A5  
## 1451 1678 40 NA B3  
## 2870 6170 78 NA A9  
## 2937 5561 52 NA B2  
## 3155 5562 4 NA C0  
## 5631 3541 138 NA A5  
## address apart.ment.number zip.code  
## 1201 33 HATTIE JONES COURT 11213  
## 1451 568 MAC DONOUGH STREET 11233  
## 2870 1627 71ST STREET 11204  
## 2937 1865 67TH STREET 11204  
## 3155 6613 19TH AVENUE 11204  
## 5631 216 BRISTOL STREET 11212  
## residential.units commercial.units total.units land.square.feet  
## 1201 1 0 1 2,173  
## 1451 2 0 2 1,800  
## 2870 1 0 1 1,742  
## 2937 2 0 2 3,000  
## 3155 3 0 3 3,000  
## 5631 1 0 1 1,800  
## gross.square.feet year.built tax.class.at.time.of.sale  
## 1201 1,452 1985 1  
## 1451 2,592 1899 1  
## 2870 1,569 1935 1  
## 2937 2,064 1920 1  
## 3155 2,394 1930 1  
## 5631 1,148 1986 1  
## building.class.at.time.of.sale sale.price sale.date sale.price.n  
## 1201 A5 $1 2012-08-24 1  
## 1451 B3 $1 2013-02-08 1  
## 2870 A9 $1 2012-11-05 1  
## 2937 B2 $1 2012-12-19 1  
## 3155 C0 $1 2013-04-18 1  
## 5631 A5 $1 2012-10-25 1  
## gross.sqft land.sqft  
## 1201 1452 2173  
## 1451 2592 1800  
## 2870 1569 1742  
## 2937 2064 3000  
## 3155 2394 3000  
## 5631 1148 1800

# 실제 판매가 아닌것 같은 이상값 제거  
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))



## 연습문제

맨하탄 지역에 대한 rollingsales\_manhattan.xls를 불러와서 정제하고 동네와 시간에 따라 데이터를 시각화하여 비교하시오.