Cleaning up data




Источники проблем в данных

- Особенности формата (лишние строки в начале файла, наличие/отсутствие заголовка, нетрадиционные разделители, etc.)
- Отсутствие некоторых данных (na)
- Типы данных (перевод строк в числа и т.п.)
- Выбросы, которые искажают общий тренд

Данные про жилье

```
require(gdata)

bk <- read.xls("rollingsales_brooklyn.xls",pattern="BOROUGH")
#все что до строки, содержащей , "BOROUGH", не читаем


head(bk) #смотрим на данные
summary(bk) #сводная статистика, чего сколько
```

Чистим данные

```
head(bk$SALE.PRICE)
[1] $403,572 $218,010 $952,311 $842,692 $815,288 $815,288
3318
Levels: $0 $1 $10 $100 $1,000 $10,000 $100,000 $1,000,000 ...
$999,999
```

Переводим цены в числовой формат

```
>bk$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","", bk
$SALE.PRICE))
# убираем все кроме цифр, т.е. заменяем все кроме цифр
на “”
```

Чистим данные

Смотрим, для скольки объектов у нас нет данных про цены

```
>count(is.na(bk$SALE.PRICE.N))#sum
```

Сделаем все имена столбцов маленькими буквами

```
>names(bk) <- tolower(names(bk))
```

Приведем в порядок площади

```
>bk$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", bk
$gross.square.feet))
>bk$land.sqft <- as.numeric(gsub("[^[:digit:]]","", bk
$land.square.feet))
```

Приведем в порядок даты

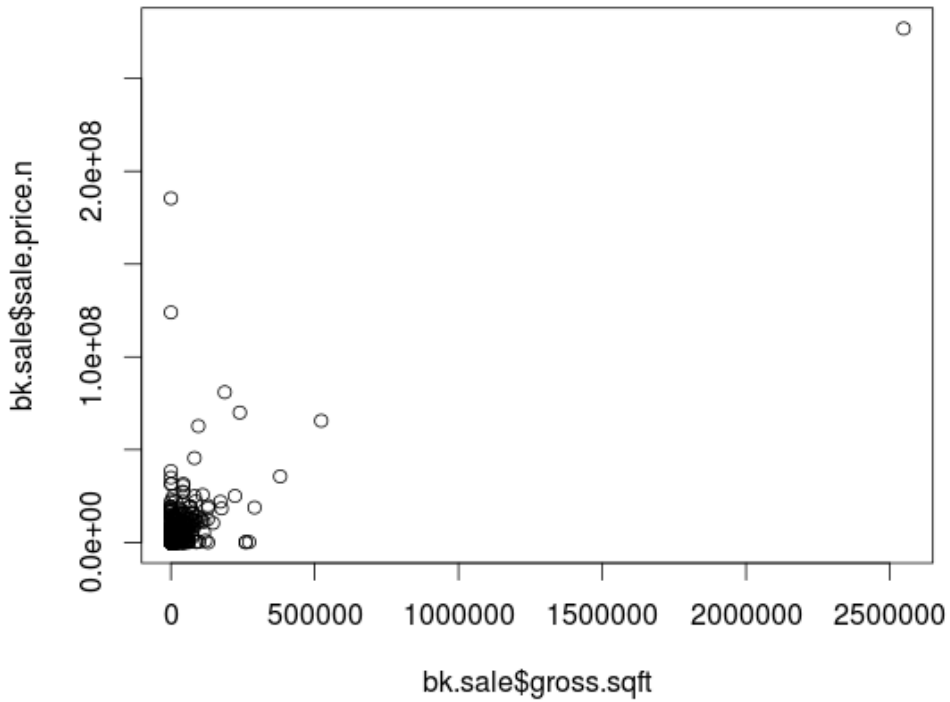
```
>bk$sale.date <- as.Date(bk$sale.date)
>bk$year.built <- as.numeric(as.character(bk$year.built))
```

Надоело писать длинные имена?
Работаем с одной таблицей?
Нет проблем!

```
>attach(bk)#теперь по-умолчанию работаем только с bk
>hist(sale.price.n)#обращаемся прямо по имени поля
>hist(sale.price.n[sale.price.n>0])
>hist(gross.sqft[sale.price.n==0])
>detach(bk)#закончили работать, открепляемся!
```

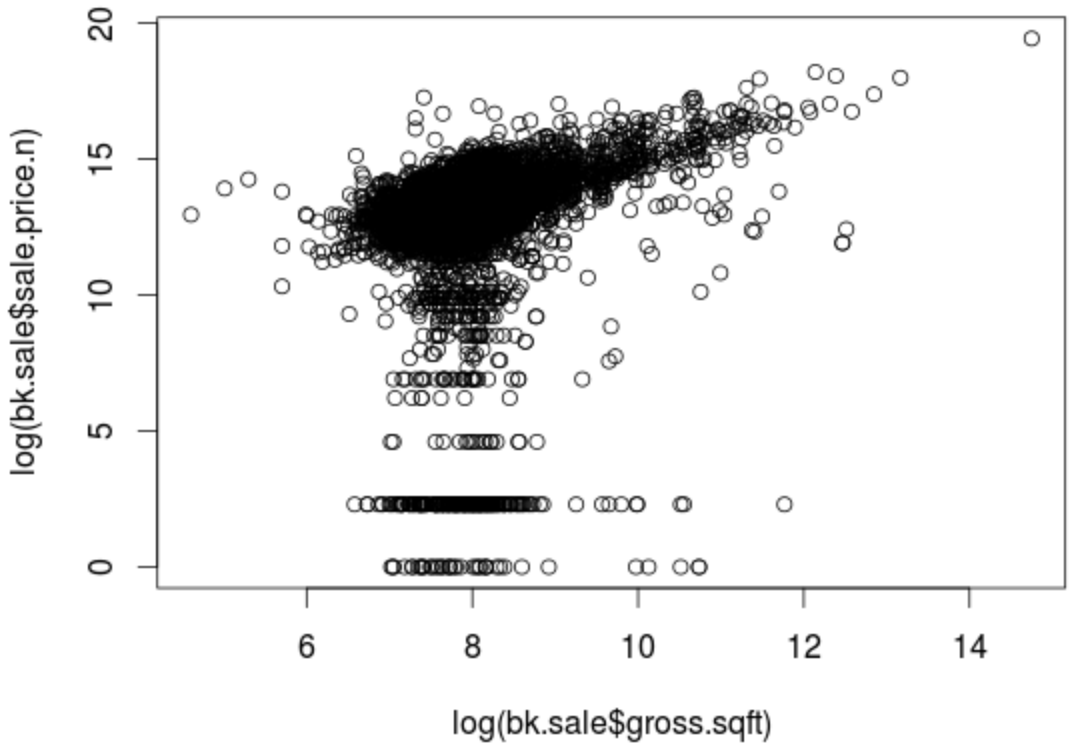
Теперь беглый анализ, как устроены данные

```
>bk.sale <- bk[bk$sale.price.n!=0,]
>plot(bk.sale$gross.sqft,bk.sale
$sale.price.n)
```



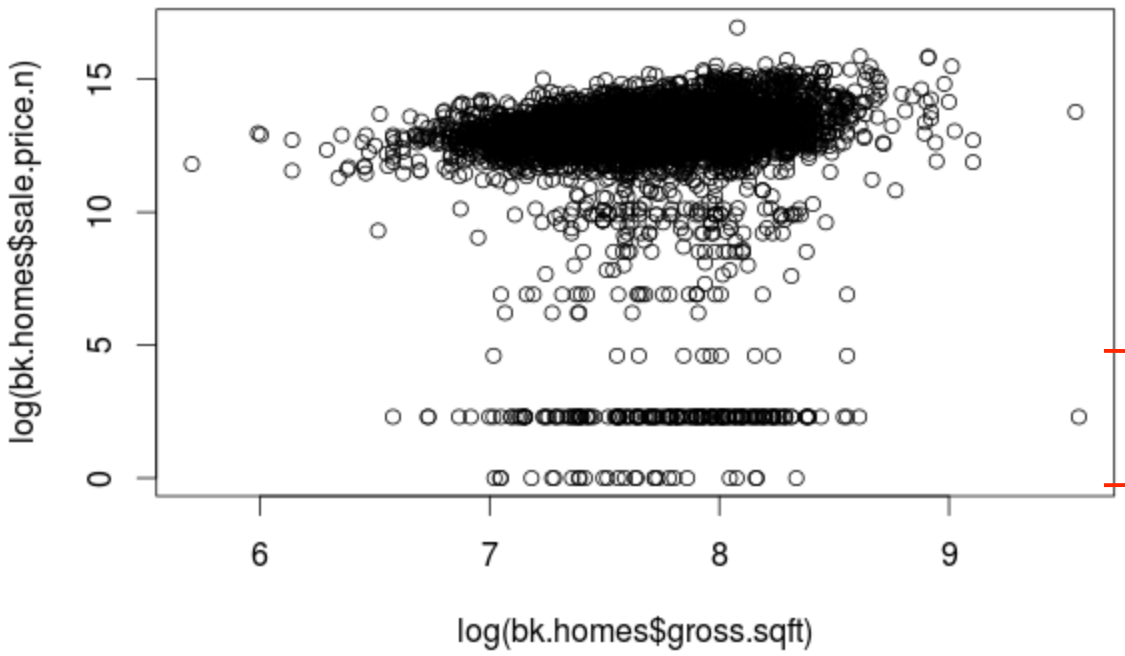
“Вынем” данные из нуля

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



Выберем для анализа только дома (категория содержит в названии “FAMILY”)

```
>bk.homes <- bk.sale[which(grepl("FAMILY", bk.sale
$building.class.category)),]
>plot(log(bk.homes$gross.sqft),log(bk.homes$sale.price.n))
```



Похоже, эти дома какие-то особенные

Уберем “особенные” дома

```
>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))
```