BronxHousingData

Chris Woodard, Nathan Mowat, Saloman Gilles, Jessica Wheeler

June 1, 2016

We used the [Rolling Housing Sales for NYC](http://www1.nyc.gov/home/search/index.page?search-terms=Rolling+sales+update) website and looked at the Bronx dataset for analysis.

* The first steps towards analysis are loading the datafile and then cleaning the datafile. Once the data is cleaned, we were able to analyze some data on housing trends in the Bronx.

### Getting Started - Loading the datafile:

## @knitr dataLoader  
  
#this file contains coed for reading in the data to analyse  
#---------------------------------------------------------------------------------------  
# this file contains code for for reading in the data file for analysis --  
#---------------------------------------------------------------------------------------  
### ------------------------------------------------------------------------------------  
### NOTE : must be loaded first --  
### ------------------------------------------------------------------------------------  
  
#STEP 1: DOWNLOAD PACKAGES AND LOAD THEM.  
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

require(plyr)

## Loading required package: plyr

#STEP 2: GET DATA INTO R FROM DROPBOX  
setwd("/Users/JessicaSibal/Dropbox/Wk3Bronx")  
bx <- read.csv("/Users/JessicaSibal/Dropbox/Wk3Bronx/Data/rollingsales\_bronx.csv",skip=4,header=TRUE)

#STEP 3: CHECK DATA  
head(bx)  
summary(bx)  
str(bx)

### Cleaning the Bronx datafile:

## @knitr dataCleaner  
  
#---------------------------------------------------------------------------------------  
# this file contains code for cleaning the data to analyse --  
#---------------------------------------------------------------------------------------  
### ------------------------------------------------------------------------------------  
### NOTE : must load after LoadDataFile.R before it works after necessary libraries --  
### ------------------------------------------------------------------------------------  
  
#STEP 4: CLEAN DATA  
bx$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","", bx$SALE.PRICE)) #Changes SALE.PRICE to numeric, and cleans it up creating a new variable  
count(is.na(bx$SALE.PRICE.N)) #Counts all the NA's in the dataset under SALE.PRICE.N

## x freq  
## 1 FALSE 4856  
## 2 TRUE 15257

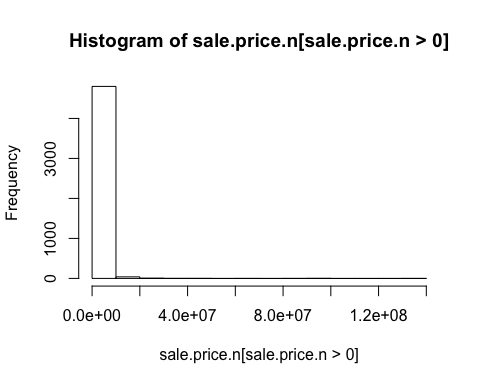
names(bx) <- tolower(names(bx)) #Changes all column names to lower case  
bx$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", bx$gross.square.feet)) #Gets rid of leading digits  
bx$land.sqft <- as.numeric(gsub("[^[:digit:]]","", bx$land.square.feet)) #Gets rid of leading digits  
bx$sale.date <- as.Date(bx$sale.date) # Changes sale.date to date format in R  
bx$year.built <- as.numeric(as.character(bx$year.built)) #Changes variable to numeric from factor

### Bronx Data Analysis:

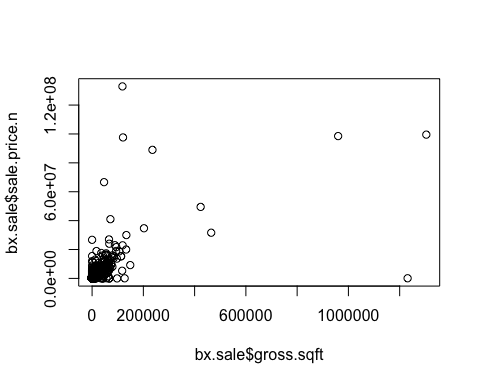
## @knitr dataAnalyser  
  
#---------------------------------------------------------------------------------------  
# this file contains analysis, includes histograms, plots, summary --   
#---------------------------------------------------------------------------------------  
### ------------------------------------------------------------------------------------  
### NOTE : must load after cleanDataFile.R before it works after necessary libraries --  
### ------------------------------------------------------------------------------------  
  
#STEP 5: VISUALIZE DATA OF SALES AND GROSS.SQFT  
attach(bx)  
hist(sale.price.n) # regular histogram of the data, doesn't look right.



hist(sale.price.n[sale.price.n>0]) #noticed over 15,000 missing data (or NA's) for sales.price.n



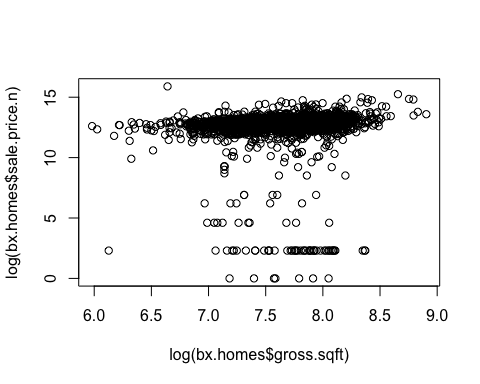
detach(bx)  
  
bx.sale <- bx[bx$sale.price.n!=0,] #Keeping only the actual sales: DOESNT GET RID OF NA's, NOT ZEROES IN DATA SET, THEY ARE DASHES (-)  
plot(bx.sale$gross.sqft,bx.sale$sale.price.n) #ScatterPlot of gross.sqft (on x-axis) vs sale.price.n (on y-axis)



#STEP 6: LOOK AT 1-3 FAMILY HOMES  
bx.homes <- bx.sale[which(grepl("FAMILY",bx.sale$building.class.category)),]   
dim(bx.homes) # shows 2589 Rows and 24 Columns

## [1] 2589 24

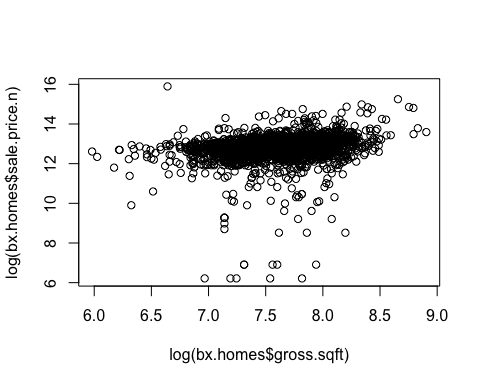
plot(log(bx.homes$gross.sqft),log(bx.homes$sale.price.n)) #Plot of Log transform of gross.sqft and sale.price.n



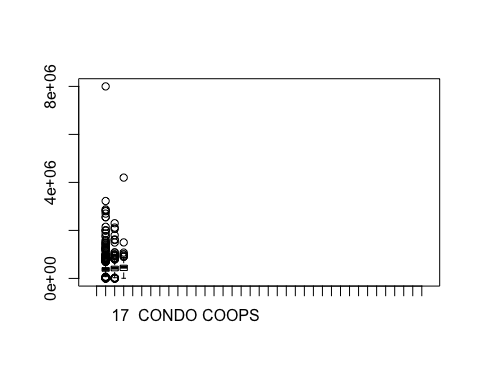
summary(bx.homes[which(bx.homes$sale.price.n<100000),]) #Summary of sale.price.n below $100,000

## borough neighborhood  
## Min. :2 BAYCHESTER :14   
## 1st Qu.:2 MORRISANIA/LONGWOOD :12   
## Median :2 SOUNDVIEW :12   
## Mean :2 WAKEFIELD : 8   
## 3rd Qu.:2 BRONXDALE : 7   
## Max. :2 THROGS NECK : 7   
## (Other) :68   
## building.class.category  
## 02 TWO FAMILY DWELLINGS :62   
## 01 ONE FAMILY DWELLINGS :42   
## 03 THREE FAMILY DWELLINGS :24   
## : 0   
## 04 TAX CLASS 1 CONDOS : 0   
## 05 TAX CLASS 1 VACANT LAND : 0   
## (Other) : 0   
## tax.class.at.present block lot ease.ment   
## 1 :128 Min. :2278 Min. : 1.00 Mode:logical   
## : 0 1st Qu.:3226 1st Qu.: 20.75 NA's:128   
## 1A : 0 Median :4072 Median : 46.00   
## 1B : 0 Mean :4113 Mean : 92.44   
## 1D : 0 3rd Qu.:4980 3rd Qu.: 82.50   
## 2 : 0 Max. :5870 Max. :2294.00   
## (Other): 0   
## building.class.at.present address   
## B1 :30 1281 UNION AVE : 4   
## C0 :24 1774 CROSS BRONX EXPRESSW : 2   
## A1 :19 3230 JOHNSON AVENUE : 2   
## B2 :14 4465 PARK AVE : 2   
## B3 :13 1010 QUINCY AVENUE : 1   
## A5 : 8 1041 EAST 223 STREET : 1   
## (Other):20 (Other) :116   
## apartment.number zip.code residential.units commercial.units   
## :128 Min. :10453 Min. :1.000 Min. :0.00000   
## : 0 1st Qu.:10459 1st Qu.:1.000 1st Qu.:0.00000   
## 1 : 0 Median :10465 Median :2.000 Median :0.00000   
## 1-Jan : 0 Mean :10464 Mean :1.859 Mean :0.03125   
## 1-Mar : 0 3rd Qu.:10469 3rd Qu.:2.000 3rd Qu.:0.00000   
## 1-Nov : 0 Max. :10475 Max. :3.000 Max. :1.00000   
## (Other) : 0   
## total.units land.square.feet gross.square.feet year.built   
## Min. :1.000 Min. : 297 Min. : 458 Min. :1800   
## 1st Qu.:1.000 1st Qu.: 1939 1st Qu.:1479 1st Qu.:1910   
## Median :2.000 Median : 2500 Median :2150 Median :1928   
## Mean :1.891 Mean : 2651 Mean :2137 Mean :1933   
## 3rd Qu.:2.000 3rd Qu.: 2889 3rd Qu.:2709 3rd Qu.:1950   
## Max. :3.000 Max. :10690 Max. :4320 Max. :2012   
##   
## tax.class.at.time.of.sale building.class.at.time.of.sale sale.price  
## Min. :1 B1 :30 $10 :47   
## 1st Qu.:1 C0 :24 $100 :10   
## Median :1 A1 :19 $1 : 7   
## Mean :1 B2 :14 $1,000 : 5   
## 3rd Qu.:1 B3 :13 $500 : 5   
## Max. :1 A5 : 8 $50,000 : 4   
## (Other):20 (Other) :50   
## sale.date sale.price.n gross.sqft land.sqft   
## Min. :0003-03-16 Min. : 1 Min. : 458 Min. : 297   
## 1st Qu.:0005-08-15 1st Qu.: 10 1st Qu.:1479 1st Qu.: 1939   
## Median :0007-09-15 Median : 300 Median :2150 Median : 2500   
## Mean :0007-11-14 Mean :19036 Mean :2137 Mean : 2651   
## 3rd Qu.:0009-08-15 3rd Qu.:30000 3rd Qu.:2709 3rd Qu.: 2889   
## Max. :0012-11-15 Max. :97000 Max. :4320 Max. :10690   
## NA's :85

#STEP 7: REMOVE OUTLIERS THAT DON'T SEEM LIKE ACTUAL SALES  
bx.homes$outliers <- (log(bx.homes$sale.price.n) <=5) + 0 #Removed the log data that was less than 5  
bx.homes <- bx.homes[which(bx.homes$outliers==0),]   
plot(log(bx.homes$gross.sqft),log(bx.homes$sale.price.n)) #ScatterPlot of Log gross.sqft vs sale.price.n with outliers removed



# Based on the log plot of gross square feet and sale price, it seems that Gross square footage is not a good indicator of sale price based on the spread of the data.  
  
plot(bx.homes$building.class.category, bx.homes$sale.price.n)



#The 1 family homes have less variance as can be seen by the boxplot but they have a lot of outliers as well. The 1-family homes could encompass low or wealthy areas in cities which could account for all of the outliers observed. In the 2-family and 3-family homes, there is a bit more variance as the spread of the data is larger.