Week6Assignment

Ashwin Thota

June 18, 2017

# read csv file  
bk <- read.csv("https://raw.githubusercontent.com/AshwinThotaDS/Week6/master/Data/rollingsales\_statenisland.csv",skip=4,header=TRUE)

## Check the data  
head(bk)

## BOROUGH NEIGHBORHOOD  
## 1 5 ANNADALE   
## 2 5 ANNADALE   
## 3 5 ANNADALE   
## 4 5 ANNADALE   
## 5 5 ANNADALE   
## 6 5 ANNADALE   
## BUILDING.CLASS.CATEGORY TAX.CLASS.AT.PRESENT BLOCK  
## 1 01 ONE FAMILY DWELLINGS 1 5391  
## 2 01 ONE FAMILY DWELLINGS 1 5397  
## 3 01 ONE FAMILY DWELLINGS 1 5408  
## 4 01 ONE FAMILY DWELLINGS 1 6205  
## 5 01 ONE FAMILY DWELLINGS 1 6209  
## 6 01 ONE FAMILY DWELLINGS 1 6211  
## LOT EASE.MENT BUILDING.CLASS.AT.PRESENT  
## 1 30 NA A1  
## 2 60 NA A1  
## 3 7 NA A2  
## 4 24 NA A5  
## 5 69 NA A5  
## 6 4 NA A5  
## ADDRESS APARTMENT.NUMBER ZIP.CODE  
## 1 1306 ARDEN AVENUE 10312  
## 2 33 EYLANDT STREET 10312  
## 3 192 BATHGATE STREET 10312  
## 4 79 EAGAN AVENUE 10312  
## 5 195 MOSELY AVENUE 10312  
## 6 35 EAGAN AVENUE 10312  
## RESIDENTIAL.UNITS COMMERCIAL.UNITS TOTAL.UNITS LAND.SQUARE.FEET  
## 1 1 - 1 7,500   
## 2 1 - 1 10,120   
## 3 1 - 1 2,500   
## 4 1 - 1 1,845   
## 5 1 - 1 2,665   
## 6 1 - 1 1,508   
## GROSS.SQUARE.FEET YEAR.BUILT TAX.CLASS.AT.TIME.OF.SALE  
## 1 1,890 1952 1  
## 2 3,516 1984 1  
## 3 572 1930 1  
## 4 1,854 1986 1  
## 5 1,592 2007 1  
## 6 1,521 1986 1  
## BUILDING.CLASS.AT.TIME.OF.SALE SALE.PRICE SALE.DATE  
## 1 A1 - 5/4/2016  
## 2 A1 - 8/10/2016  
## 3 A2 290,000 9/2/2016  
## 4 A5 425,000 7/29/2016  
## 5 A5 579,000 11/21/2016  
## 6 A5 - 7/25/2016

summary(bk)

## BOROUGH NEIGHBORHOOD   
## Min. :5 GREAT KILLS : 703   
## 1st Qu.:5 NEW SPRINGVILLE : 465   
## Median :5 BULLS HEAD : 359   
## Mean :5 MIDLAND BEACH : 347   
## 3rd Qu.:5 ELTINGVILLE : 331   
## Max. :5 WEST NEW BRIGHTON : 300   
## (Other) :5576   
## BUILDING.CLASS.CATEGORY  
## 01 ONE FAMILY DWELLINGS :4579   
## 02 TWO FAMILY DWELLINGS :1848   
## 04 TAX CLASS 1 CONDOS : 479   
## 05 TAX CLASS 1 VACANT LAND : 400   
## 13 CONDOS - ELEVATOR APARTMENTS : 166   
## 31 COMMERCIAL VACANT LAND : 86   
## (Other) : 523   
## TAX.CLASS.AT.PRESENT BLOCK LOT EASE.MENT   
## 1 :6531 Min. : 1 Min. : 1.0 Mode:logical   
## 1A : 469 1st Qu.:1116 1st Qu.: 23.0 NA's:8081   
## 1B : 379 Median :3100 Median : 52.0   
## 4 : 315 Mean :3278 Mean : 183.1   
## 2 : 310 3rd Qu.:5317 3rd Qu.: 125.0   
## 2A : 46 Max. :8050 Max. :5357.0   
## (Other): 31   
## BUILDING.CLASS.AT.PRESENT  
## A5 :1850   
## A1 :1286   
## B2 :1207   
## A2 : 930   
## R3 : 469   
## V0 : 377   
## (Other):1962   
## ADDRESS APARTMENT.NUMBER  
## 100 COLFAX AVENUE : 22 :7666   
## 145 LINCOLN AVENUE : 21 A : 10   
## 2 ELMWOOD PARK DRIVE : 19 1 : 9   
## 755 NARROWS ROAD NORTH : 12 1B : 9   
## 5 WINDHAM LOOP : 10 2A : 8   
## 1235 FOREST HILL ROAD : 8 2B : 8   
## (Other) :7989 (Other) : 371   
## ZIP.CODE RESIDENTIAL.UNITS COMMERCIAL.UNITS TOTAL.UNITS   
## Min. : 0 1 :4893 0 :7281 1 :4976   
## 1st Qu.:10304 2 :1763 - : 523 2 :1783   
## Median :10307 0 : 769 1 : 200 0 : 597   
## Mean :10032 1 : 386 2 : 37 1 : 390   
## 3rd Qu.:10312 2 : 110 3 : 16 3 : 109   
## Max. :10314 3 : 63 1 : 10 2 : 108   
## (Other): 97 (Other): 14 (Other): 118   
## LAND.SQUARE.FEET GROSS.SQUARE.FEET YEAR.BUILT   
## - : 796 - :1312 Min. : 0   
## 4,000 : 538 1,200 : 124 1st Qu.:1935   
## 2,500 : 272 1,800 : 89 Median :1970   
## 5,000 : 234 1,600 : 77 Mean :1849   
## 6,000 : 141 1,440 : 70 3rd Qu.:1990   
## 2,000 : 119 1,400 : 60 Max. :2016   
## (Other):5981 (Other):6349   
## TAX.CLASS.AT.TIME.OF.SALE BUILDING.CLASS.AT.TIME.OF.SALE SALE.PRICE   
## Min. :1.000 A5 :1826 - :2178   
## 1st Qu.:1.000 A1 :1280 10 : 110   
## Median :1.000 B2 :1193 450,000 : 70   
## Mean :1.163 A2 : 935 550,000 : 64   
## 3rd Qu.:1.000 R3 : 476 350,000 : 62   
## Max. :4.000 V0 : 397 300,000 : 57   
## (Other):1974 (Other) :5540   
## SALE.DATE   
## 3/1/2017 : 89   
## 6/29/2016: 66   
## 8/31/2016: 65   
## 6/30/2016: 64   
## 6/16/2016: 62   
## 8/26/2016: 54   
## (Other) :7681

str(bk) # Very handy function!

## 'data.frame': 8081 obs. of 21 variables:  
## $ BOROUGH : int 5 5 5 5 5 5 5 5 5 5 ...  
## $ NEIGHBORHOOD : Factor w/ 58 levels "ANNADALE ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ BUILDING.CLASS.CATEGORY : Factor w/ 29 levels "01 ONE FAMILY DWELLINGS ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ TAX.CLASS.AT.PRESENT : Factor w/ 10 levels " ","1","1A",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ BLOCK : int 5391 5397 5408 6205 6209 6211 6211 6212 6212 6212 ...  
## $ LOT : int 30 60 7 24 69 4 15 47 50 59 ...  
## $ EASE.MENT : logi NA NA NA NA NA NA ...  
## $ BUILDING.CLASS.AT.PRESENT : Factor w/ 81 levels " ","A0","A1",..: 3 3 4 6 6 6 6 6 6 6 ...  
## $ ADDRESS : Factor w/ 7459 levels "1 BOLIVAR STREET ",..: 807 3740 1986 6565 2011 3954 1544 4628 4082 6975 ...  
## $ APARTMENT.NUMBER : Factor w/ 278 levels " ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ ZIP.CODE : int 10312 10312 10312 10312 10312 10312 10312 10312 10312 10312 ...  
## $ RESIDENTIAL.UNITS : Factor w/ 20 levels " - "," 1 ",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ COMMERCIAL.UNITS : Factor w/ 13 levels " - "," 1 ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ TOTAL.UNITS : Factor w/ 25 levels " - "," 1 ",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ LAND.SQUARE.FEET : Factor w/ 2359 levels " - "," 1,000 ",..: 2144 341 773 282 852 169 169 865 938 219 ...  
## $ GROSS.SQUARE.FEET : Factor w/ 1693 levels " - "," 1,000 ",..: 576 1284 1495 553 387 336 336 320 320 356 ...  
## $ YEAR.BUILT : int 1952 1984 1930 1986 2007 1986 1986 2000 2000 2005 ...  
## $ TAX.CLASS.AT.TIME.OF.SALE : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ BUILDING.CLASS.AT.TIME.OF.SALE: Factor w/ 81 levels "A0","A1","A2",..: 2 2 3 5 5 5 5 5 5 5 ...  
## $ SALE.PRICE : Factor w/ 1570 levels " - "," 1 ",..: 1 1 443 794 1134 1 638 1 909 900 ...  
## $ SALE.DATE : Factor w/ 311 levels "1/1/2017","1/10/2017",..: 197 256 294 247 69 243 34 38 199 309 ...

#Compactly display the internal structure of an R object.

## clean/format the data with regular expressions  
## More on these later. For now, know that the  
## pattern "[^[:digit:]]" refers to members of the variable name that  
## start with digits. We use the gsub command to replace them with a blank space.  
# We create a new variable that is a "clean' version of sale.price.  
# And sale.price.n is numeric, not a factor.  
library(plyr)

## Warning: package 'plyr' was built under R version 3.3.3

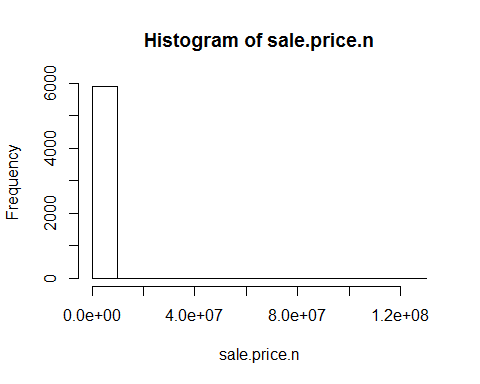
bk$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","", bk$SALE.PRICE))  
count(is.na(bk$SALE.PRICE.N))

## x freq  
## 1 FALSE 5903  
## 2 TRUE 2178

names(bk) <- tolower(names(bk)) # make all variable names lower case

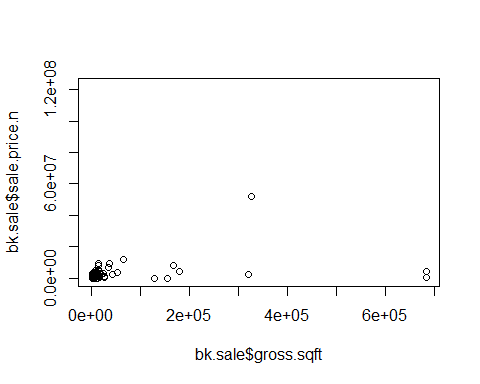
bk$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", bk$gross.square.feet))  
bk$land.sqft <- as.numeric(gsub("[^[:digit:]]","", bk$land.square.feet))  
bk$year.built <- as.numeric(as.character(bk$year.built))

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



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(log10(bk.sale$gross.sqft),log10(bk.sale$sale.price.n))

