Rolling Sales Manhattan-visualizations - Word doc

## Data Load

Data is loaded from a local copy of Rolling Sales data for Manhattan sourced at "analysis/data/rollingsales\_manhattan.csv". This is done by calling the load script "analysis/Load.R". The structure of the loaded data set is shown below.

source("analysis/Load.R",echo = TRUE)

##   
## > readURL <- "analysis/data/rollingsales\_manhattan.csv"  
##   
## > bk <- read.csv(readURL, skip = 4, header = TRUE)  
##   
## > str(bk)  
## 'data.frame': 23412 obs. of 21 variables:  
## $ BOROUGH : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ NEIGHBORHOOD : Factor w/ 39 levels "ALPHABET CITY ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ BUILDING.CLASS.CATEGORY : Factor w/ 42 levels "01 ONE FAMILY DWELLINGS ",..: 3 3 4 5 5 5 5 5 5 5 ...  
## $ TAX.CLASS.AT.PRESENT : Factor w/ 8 levels "1","1A","1C",..: 1 1 3 6 6 4 4 4 4 4 ...  
## $ BLOCK : int 376 377 399 373 373 377 377 385 387 387 ...  
## $ LOT : int 24 66 1101 16 17 2 2 2 119 119 ...  
## $ EASE.MENT : logi NA NA NA NA NA NA ...  
## $ BUILDING.CLASS.AT.PRESENT : Factor w/ 123 levels "A1","A4","A5",..: 10 10 92 11 11 17 17 17 17 17 ...  
## $ ADDRESS : Factor w/ 12053 levels "1 5TH AVENUE, 11F/G ",..: 5794 5235 5071 7252 7266 738 738 4096 10862 10862 ...  
## $ APARTMENT.NUMBER : Factor w/ 2758 levels "- ",..: 2 2 1 2 2 2 2 2 2 2 ...  
## $ ZIP.CODE : int 10009 10009 10009 10009 10009 10009 10009 10009 10009 10009 ...  
## $ RESIDENTIAL.UNITS : int 3 3 1 10 10 22 22 12 22 22 ...  
## $ COMMERCIAL.UNITS : int 0 0 0 0 0 3 3 4 2 2 ...  
## $ TOTAL.UNITS : int 3 3 1 10 10 25 25 16 24 24 ...  
## $ LAND.SQUARE.FEET : int 2059 2381 0 2204 2204 4510 4510 4186 3445 3445 ...  
## $ GROSS.SQUARE.FEET : int 3696 3084 0 8625 8625 19830 19830 10588 13546 13546 ...  
## $ YEAR.BUILT : int 1900 1899 1955 1899 1900 1900 1900 1900 1920 1920 ...  
## $ TAX.CLASS.AT.TIME.OF.SALE : int 1 1 1 2 2 2 2 2 2 2 ...  
## $ BUILDING.CLASS.AT.TIME.OF.SALE: Factor w/ 123 levels "A1","A4","A5",..: 10 10 92 11 11 17 17 17 17 17 ...  
## $ SALE.PRICE : Factor w/ 6619 levels " $- "," $1 ",..: 2989 2312 5042 2357 1 1579 1 1 1160 1 ...  
## $ SALE.DATE : Factor w/ 339 levels "1/1/2016","1/11/2016",..: 40 236 227 65 65 311 311 143 115 115 ...

## Data Conversion and Cleanup

The Rolling Sales data is now converted and cleaned up using the script "analysis/DataConversion.R". The sales variables are changed to account for bad data and filtered for actual sales of 1-, 2-, nd 3- family homes. The new structure of the converted data set is shown below.

source("analysis/DataConversion.R",echo = TRUE)

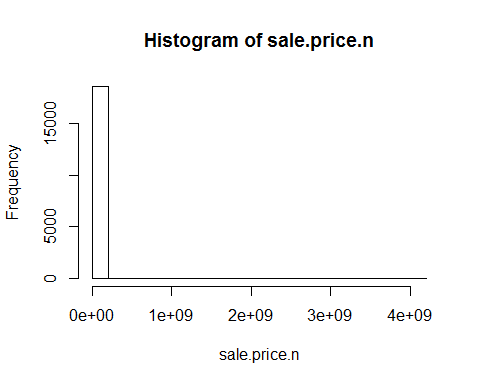
##   
## > bk$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]",   
## + "", bk$SALE.PRICE))  
##   
## > count(is.na(bk$SALE.PRICE.N))  
## x freq  
## 1 FALSE 18611  
## 2 TRUE 4801  
##   
## > names(bk) <- tolower(names(bk))  
##   
## > bk$gross.sqft <- as.numeric(gsub("[^[:digit:]]", "",   
## + bk$gross.square.feet))  
##   
## > bk$land.sqft <- as.numeric(gsub("[^[:digit:]]", "",   
## + bk$land.square.feet))  
##   
## > bk$sale.date <- as.Date(bk$sale.date, "%m/%d/%Y")  
##   
## > bk$year.built <- as.numeric((bk$year.built))  
##   
## > str(bk)  
## 'data.frame': 23412 obs. of 24 variables:  
## $ borough : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ neighborhood : Factor w/ 39 levels "ALPHABET CITY ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ building.class.category : Factor w/ 42 levels "01 ONE FAMILY DWELLINGS ",..: 3 3 4 5 5 5 5 5 5 5 ...  
## $ tax.class.at.present : Factor w/ 8 levels "1","1A","1C",..: 1 1 3 6 6 4 4 4 4 4 ...  
## $ block : int 376 377 399 373 373 377 377 385 387 387 ...  
## $ lot : int 24 66 1101 16 17 2 2 2 119 119 ...  
## $ ease.ment : logi NA NA NA NA NA NA ...  
## $ building.class.at.present : Factor w/ 123 levels "A1","A4","A5",..: 10 10 92 11 11 17 17 17 17 17 ...  
## $ address : Factor w/ 12053 levels "1 5TH AVENUE, 11F/G ",..: 5794 5235 5071 7252 7266 738 738 4096 10862 10862 ...  
## $ apartment.number : Factor w/ 2758 levels "- ",..: 2 2 1 2 2 2 2 2 2 2 ...  
## $ zip.code : int 10009 10009 10009 10009 10009 10009 10009 10009 10009 10009 ...  
## $ residential.units : int 3 3 1 10 10 22 22 12 22 22 ...  
## $ commercial.units : int 0 0 0 0 0 3 3 4 2 2 ...  
## $ total.units : int 3 3 1 10 10 25 25 16 24 24 ...  
## $ land.square.feet : int 2059 2381 0 2204 2204 4510 4510 4186 3445 3445 ...  
## $ gross.square.feet : int 3696 3084 0 8625 8625 19830 19830 10588 13546 13546 ...  
## $ year.built : num 1900 1899 1955 1899 1900 ...  
## $ tax.class.at.time.of.sale : int 1 1 1 2 2 2 2 2 2 2 ...  
## $ building.class.at.time.of.sale: Factor w/ 123 levels "A1","A4","A5",..: 10 10 92 11 11 17 17 17 17 17 ...  
## $ sale.price : Factor w/ 6619 levels " $- "," $1 ",..: 2989 2312 5042 2357 1 1579 1 1 1160 1 ...  
## $ sale.date : Date, format: "2015-10-22" "2015-06-24" ...  
## $ sale.price.n : num 3775000 2900000 6995000 20000000 NA ...  
## $ gross.sqft : num 3696 3084 0 8625 8625 ...  
## $ land.sqft : num 2059 2381 0 2204 2204 ...

## Sales Price Data

We are now able to see some trends in the data. Using the script "analysis/Analysis1.R" we can generate a histogram for the sales price data.

source("analysis/Analysis1.R",echo = TRUE)

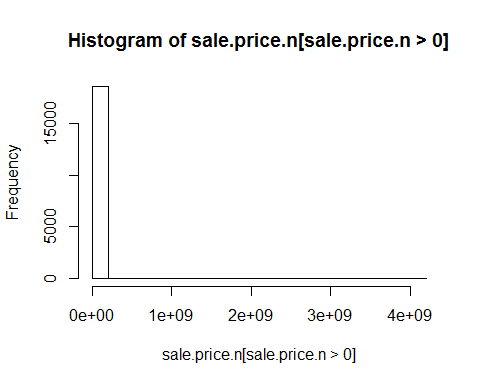
##   
## > attach(bk)  
##   
## > hist(sale.price.n)



Using the script "analysis/Analysis2.R" we can generate a histogram for the sales price data for all sales prices greater than $0. This plot looks the same as the prior one.

source("analysis/Analysis2.R",echo = TRUE)

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



Using the script "analysis/Analysis3.R" we can generate a summary for the sale price data.

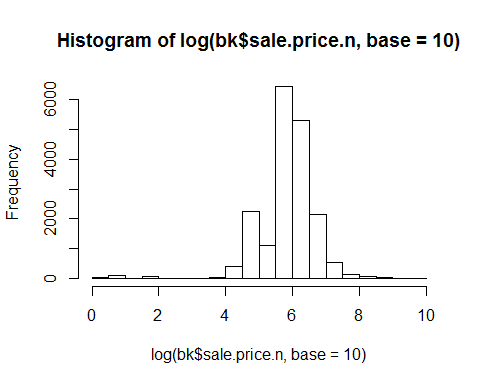
source("analysis/Analysis3.R",echo = TRUE)

##   
## > summary(sale.price.n)  
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.000e+00 3.950e+05 8.550e+05 3.456e+06 2.000e+06 4.111e+09 4801

## Conversion of Sales Price Data to Log Scale

The prior plots revealed that the data was heavily concentrated therefore, we took the log of the same data. Using the script "analysis/Analysis4.R" we can generate a histogram for the sales price data on a log scale (base 10). We see that the data appears to not have a strong log form, but tend to move in a positive diration aside from outliers.

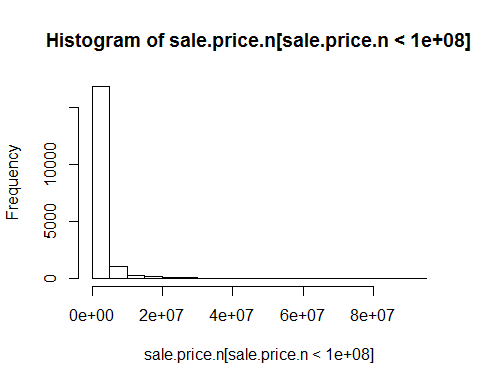
source("analysis/Analysis4.R")



We can limit the histogram to a lesser value, since quartile 3 is $2MM. Using the script "analysis/Analysis5.R" we can generate a histogram for the sales price data on a log scale (base 10) less than 100e6.

source("analysis/Analysis5.R", echo = TRUE)

##   
## > hist(sale.price.n[sale.price.n < 1e+08])



But that still shows one big cluster. Using the script "analysis/Analysis6.R" we can see though that there are no zero's in our set.

source("analysis/Analysis6.R",echo = TRUE)

##   
## > count(sale.price.n == 0)  
## x freq  
## 1 FALSE 18611  
## 2 NA 4801  
##   
## > detach(bk)

## Further Analysis

Further Analysis can now be done. Using the script "analysis/Analysis7.R" we can get only records from the sales price data that have non-zero values. A summary of the data set is shown below.

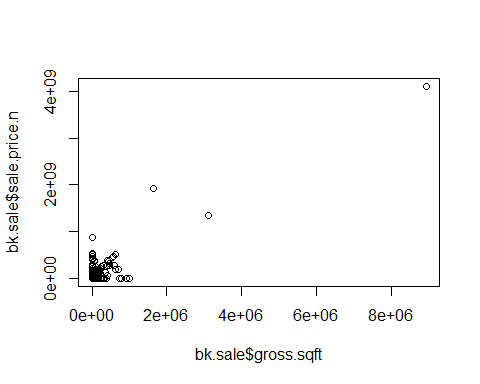
source("analysis/Analysis7.R", echo = TRUE)

##   
## > bk.sale <- bk[bk$sale.price.n != 0, ]  
##   
## > summary(bk.sale)  
## borough neighborhood   
## Min. :1 MIDTOWN WEST :3662   
## 1st Qu.:1 UPPER EAST SIDE (59-79) :1895   
## Median :1 UPPER EAST SIDE (79-96) :1467   
## Mean :1 UPPER WEST SIDE (59-79) :1341   
## 3rd Qu.:1 MIDTOWN EAST :1159   
## Max. :1 (Other) :9087   
## NA's :4801 NA's :4801   
## building.class.category  
## 13 CONDOS - ELEVATOR APARTMENTS :6099   
## 10 COOPS - ELEVATOR APARTMENTS :5860   
## 45 CONDO HOTELS :2279   
## 17 CONDO COOPS :1144   
## 09 COOPS - WALKUP APARTMENTS : 731   
## (Other) :2498   
## NA's :4801   
## tax.class.at.present block lot ease.ment   
## 2 :14023 Min. : 7 Min. : 1.0 Mode:logical   
## 4 : 3533 1st Qu.: 874 1st Qu.: 35.0 NA's:23412   
## 2C : 534 Median :1076 Median :1008.0   
## 1 : 242 Mean :1116 Mean : 793.8   
## 2B : 169 3rd Qu.:1414 3rd Qu.:1302.0   
## (Other): 110 Max. :2250 Max. :9117.0   
## NA's : 4801 NA's :4801 NA's :4801   
## building.class.at.present  
## R4 :6100   
## D4 :5633   
## RH :2279   
## R9 :1144   
## C6 : 723   
## (Other):2732   
## NA's :4801   
## address apartment.number  
## 1335 AVENUE OF THE AMERIC : 1770 :9675   
## 102 WEST 57TH STREET : 610 HU2 :1449   
## 1335 AVENUE OF THE AMER : 391 TIMES : 712   
## 551 MAIN STREET, RES : 231 5A : 77   
## 200 EAST 94TH STREET : 188 3A : 64   
## (Other) :15421 (Other) :6634   
## NA's : 4801 NA's :4801   
## zip.code residential.units commercial.units total.units   
## Min. : 0 Min. : 0.000 Min. : 0.00 Min. : 0.000   
## 1st Qu.:10016 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 0.000   
## Median :10019 Median : 0.000 Median : 0.00 Median : 1.000   
## Mean :10031 Mean : 1.984 Mean : 0.24 Mean : 2.353   
## 3rd Qu.:10028 3rd Qu.: 1.000 3rd Qu.: 0.00 3rd Qu.: 1.000   
## Max. :10463 Max. :8759.000 Max. :131.00 Max. :8805.000   
## NA's :4801 NA's :4801 NA's :4801 NA's :4801   
## land.square.feet gross.square.feet year.built   
## Min. : 0.0 Min. : 0 Min. : 0   
## 1st Qu.: 0.0 1st Qu.: 0 1st Qu.:1920   
## Median : 0.0 Median : 0 Median :1959   
## Mean : 840.6 Mean : 6841 Mean :1799   
## 3rd Qu.: 0.0 3rd Qu.: 0 3rd Qu.:1975   
## Max. :2675000.0 Max. :8942176 Max. :2015   
## NA's :4801 NA's :4801 NA's :4801   
## tax.class.at.time.of.sale building.class.at.time.of.sale  
## Min. :1.000 R4 :6099   
## 1st Qu.:2.000 D4 :5633   
## Median :2.000 RH :2279   
## Mean :2.366 R9 :1144   
## 3rd Qu.:2.000 C6 : 723   
## Max. :4.000 (Other):2733   
## NA's :4801 NA's :4801   
## sale.price sale.date sale.price.n   
## $45,000 : 217 Min. :2015-05-01 Min. :1.000e+00   
## $10 : 97 1st Qu.:2015-07-23 1st Qu.:3.950e+05   
## $850,000 : 84 Median :2015-10-08 Median :8.550e+05   
## $650,000 : 80 Mean :2015-10-14 Mean :3.456e+06   
## $1,500,000 : 76 3rd Qu.:2016-01-04 3rd Qu.:2.000e+06   
## (Other) :18057 Max. :2016-04-29 Max. :4.111e+09   
## NA's : 4801 NA's :4801 NA's :4801   
## gross.sqft land.sqft   
## Min. : 0 Min. : 0.0   
## 1st Qu.: 0 1st Qu.: 0.0   
## Median : 0 Median : 0.0   
## Mean : 6841 Mean : 840.6   
## 3rd Qu.: 0 3rd Qu.: 0.0   
## Max. :8942176 Max. :2675000.0   
## NA's :4801 NA's :4801

Using the script "analysis/Analysis8.R" we can plot Sales Prices versus Gross Square Footage.

source("analysis/Analysis8.R", echo = TRUE)

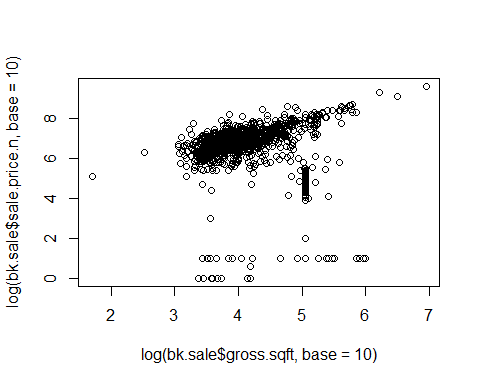
##   
## > plot(bk.sale$gross.sqft, bk.sale$sale.price.n)



Using the script "analysis/Analysis9.R" we can plot Sales Prices versus Gross Square Footage on a log scale.

source("analysis/Analysis9.R", echo = TRUE)

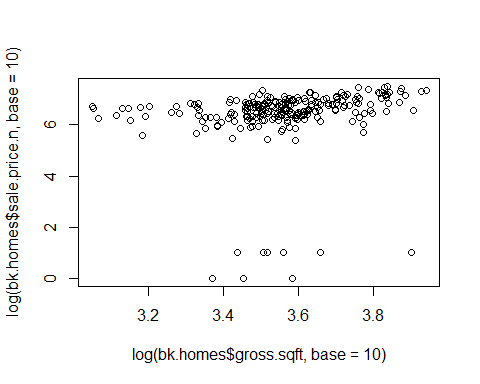
##   
## > plot(log(bk.sale$gross.sqft, base = 10), log(bk.sale$sale.price.n,   
## + base = 10))



Using the script "analysis/Analysis10.R" we can again plot Sales Prices versus Gross Square Footage but filter for homes which have family in the class type to receive a much more centered looking dataset. We can see that the variances across square footage have now become very constant and the trend looks somewhat linear. The only troubling item in the graph now is the outliers below 5 on the y-axis.

source("analysis/Analysis10.R", echo = TRUE)

##   
## > bk.homes <- bk.sale[which(grepl("FAMILY", bk.sale$building.class.category)),   
## + ]  
##   
## > dim(bk.homes)  
## [1] 242 24  
##   
## > plot(log(bk.homes$gross.sqft, base = 10), log(bk.homes$sale.price.n,   
## + base = 10))



Using the script "analysis/Analysis11.R" we can see a summary for homes with family in the class type.

source("analysis/Analysis11.R", echo = TRUE)

##   
## > summary(bk.homes[which(bk.homes$sale.price.n < 1e+05),   
## + ])  
## borough neighborhood  
## Min. :1 UPPER WEST SIDE (96-116) :3   
## 1st Qu.:1 HARLEM-EAST :2   
## Median :1 UPPER EAST SIDE (79-96) :2   
## Mean :1 GREENWICH VILLAGE-WEST :1   
## 3rd Qu.:1 KIPS BAY :1   
## Max. :1 MIDTOWN EAST :1   
## (Other) :1   
## building.class.category  
## 03 THREE FAMILY DWELLINGS :6   
## 02 TWO FAMILY DWELLINGS :3   
## 01 ONE FAMILY DWELLINGS :2   
## 04 TAX CLASS 1 CONDOS :0   
## 07 RENTALS - WALKUP APARTMENTS :0   
## 08 RENTALS - ELEVATOR APARTMENTS :0   
## (Other) :0   
## tax.class.at.present block lot ease.ment   
## 1 :11 Min. : 593 Min. : 13.00 Mode:logical   
## 1A : 0 1st Qu.:1236 1st Qu.: 38.00 NA's:11   
## 1C : 0 Median :1520 Median : 42.00   
## 2 : 0 Mean :1474 Mean : 57.55   
## 2A : 0 3rd Qu.:1849 3rd Qu.: 62.00   
## 2B : 0 Max. :1891 Max. :150.00   
## (Other): 0   
## building.class.at.present address   
## C0 :6 333 RIVERSIDE DRIVE :3   
## B1 :2 118 E 83RD :1   
## A9 :1 128 EAST 92ND STREET :1   
## B3 :1 129 WEST 78TH STREET :1   
## S1 :1 22 CHRISTOPHER STREET :1   
## A1 :0 234 EAST 49TH STREET :1   
## (Other):0 (Other) :3   
## apartment.number zip.code residential.units commercial.units   
## :11 Min. :10003 Min. :1.000 Min. :0.00000   
## - : 0 1st Qu.:10020 1st Qu.:2.000 1st Qu.:0.00000   
## (-) : 0 Median :10025 Median :3.000 Median :0.00000   
## 1 : 0 Mean :10033 Mean :2.364 Mean :0.09091   
## 1-A : 0 3rd Qu.:10032 3rd Qu.:3.000 3rd Qu.:0.00000   
## 1-Apr : 0 Max. :10128 Max. :3.000 Max. :1.00000   
## (Other) : 0   
## total.units land.square.feet gross.square.feet year.built   
## Min. :1.000 Min. :1235 Min. :2352 Min. :1899   
## 1st Qu.:2.000 1st Qu.:1572 1st Qu.:3038 1st Qu.:1900   
## Median :3.000 Median :1840 Median :3645 Median :1901   
## Mean :2.455 Mean :1876 Mean :4594 Mean :1904   
## 3rd Qu.:3.000 3rd Qu.:2292 3rd Qu.:6284 3rd Qu.:1905   
## Max. :3.000 Max. :2554 Max. :8007 Max. :1920   
##   
## tax.class.at.time.of.sale building.class.at.time.of.sale  
## Min. :1 C0 :6   
## 1st Qu.:1 B1 :2   
## Median :1 A9 :1   
## Mean :1 B3 :1   
## 3rd Qu.:1 S1 :1   
## Max. :1 A1 :0   
## (Other):0   
## sale.price sale.date sale.price.n gross.sqft   
## $10 :8 Min. :2015-05-07 Min. : 1.000 Min. :2352   
## $1 :3 1st Qu.:2015-05-27 1st Qu.: 5.500 1st Qu.:3038   
## $- :0 Median :2015-06-11 Median :10.000 Median :3645   
## $1,000 :0 Mean :2015-07-13 Mean : 7.545 Mean :4594   
## $1,000,000 :0 3rd Qu.:2015-07-26 3rd Qu.:10.000 3rd Qu.:6284   
## $1,000,391 :0 Max. :2015-12-14 Max. :10.000 Max. :8007   
## (Other) :0   
## land.sqft   
## Min. :1235   
## 1st Qu.:1572   
## Median :1840   
## Mean :1876   
## 3rd Qu.:2292   
## Max. :2554   
##