Queens\_Data\_Analysis

Bryan\_Kevin\_Ruhaab

June 15, 2017

# Queens Data Explaratory data Analysis

This documents presents the processing of cleaning and analysing queens dataset

#Required packages  
packages <- c("ggplot2","plyr", "dygraphs, dplyr")  
  
#Check if the package is installed. if not, install it.  
  
is.installed <- function(mypkg){  
 is.element(mypkg, installed.packages()[,1])  
}  
  
for(package in packages){  
 # check if package is installed  
 if (!is.installed(package)){  
 install.packages(package, repos = "https://cran.r-project.org/")  
 }else {print(paste(package, " is library already installed"))}  
}

## [1] "ggplot2 is library already installed"  
## [1] "plyr is library already installed"

## Installing package into 'C:/Users/kevolo/Documents/R/win-library/3.3'  
## (as 'lib' is unspecified)

## Warning: package 'dygraphs, dplyr' is not available (for R version 3.3.2)

#load the required packages  
print("loading the required packages!")

## [1] "loading the required packages!"

lapply(packages, require, character.only = TRUE)

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.3.3

## Loading required package: plyr

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

## Loading required package: dygraphs, dplyr

## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,  
## logical.return = TRUE, : there is no package called 'dygraphs, dplyr'

## [[1]]  
## [1] TRUE  
##   
## [[2]]  
## [1] TRUE  
##   
## [[3]]  
## [1] FALSE

sessionInfo()

## R version 3.3.2 (2016-10-31)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 10 x64 (build 14393)  
##   
## locale:  
## [1] LC\_COLLATE=English\_United States.1252   
## [2] LC\_CTYPE=English\_United States.1252   
## [3] LC\_MONETARY=English\_United States.1252  
## [4] LC\_NUMERIC=C   
## [5] LC\_TIME=English\_United States.1252   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## other attached packages:  
## [1] plyr\_1.8.4 ggplot2\_2.2.1  
##   
## loaded via a namespace (and not attached):  
## [1] Rcpp\_0.12.8 assertthat\_0.1 digest\_0.6.10 rprojroot\_1.1   
## [5] grid\_3.3.2 gtable\_0.2.0 backports\_1.0.4 magrittr\_1.5   
## [9] evaluate\_0.10 scales\_0.4.1 stringi\_1.1.2 lazyeval\_0.2.0   
## [13] rmarkdown\_1.5 tools\_3.3.2 stringr\_1.1.0 munsell\_0.4.3   
## [17] yaml\_2.1.14 colorspace\_1.3-1 htmltools\_0.3.5 knitr\_1.16   
## [21] tibble\_1.2

Loading the Queens Dataset and exploratory

# read csv file  
queens<- read.csv("C:/Users/kevolo/Desktop/smu/doing datascience/DoingDataScienceGroupProject/data/rollingsales\_queens.csv",skip=4,header=TRUE)  
  
dim(queens)

## [1] 25827 21

names(queens)

## [1] "BOROUGH" "NEIGHBORHOOD"   
## [3] "BUILDING.CLASS.CATEGORY" "TAX.CLASS.AT.PRESENT"   
## [5] "BLOCK" "LOT"   
## [7] "EASE.MENT" "BUILDING.CLASS.AT.PRESENT"   
## [9] "ADDRESS" "APARTMENT.NUMBER"   
## [11] "ZIP.CODE" "RESIDENTIAL.UNITS"   
## [13] "COMMERCIAL.UNITS" "TOTAL.UNITS"   
## [15] "LAND.SQUARE.FEET" "GROSS.SQUARE.FEET"   
## [17] "YEAR.BUILT" "TAX.CLASS.AT.TIME.OF.SALE"   
## [19] "BUILDING.CLASS.AT.TIME.OF.SALE" "SALE.PRICE"   
## [21] "SALE.DATE"

head(queens)

## BOROUGH NEIGHBORHOOD  
## 1 4 AIRPORT LA GUARDIA   
## 2 4 AIRPORT LA GUARDIA   
## 3 4 AIRPORT LA GUARDIA   
## 4 4 AIRPORT LA GUARDIA   
## 5 4 AIRPORT LA GUARDIA   
## 6 4 AIRPORT LA GUARDIA   
## BUILDING.CLASS.CATEGORY TAX.CLASS.AT.PRESENT BLOCK  
## 1 01 ONE FAMILY DWELLINGS 1 976  
## 2 01 ONE FAMILY DWELLINGS 1 976  
## 3 01 ONE FAMILY DWELLINGS 1 976  
## 4 01 ONE FAMILY DWELLINGS 1 976  
## 5 02 TWO FAMILY DWELLINGS 1 976  
## 6 03 THREE FAMILY DWELLINGS 1 949  
## LOT EASE.MENT BUILDING.CLASS.AT.PRESENT  
## 1 10 NA A5  
## 2 11 NA A5  
## 3 61 NA A5  
## 4 63 NA A5  
## 5 70 NA B1  
## 6 56 NA C0  
## ADDRESS APARTMENT.NUMBER ZIP.CODE  
## 1 21-10 81ST STREET 11370  
## 2 21-12 81ST STREET 11370  
## 3 21-21 80TH STREET 11370  
## 4 21-17 80TH STREET 11370  
## 5 21-03 80TH STREET 11370  
## 6 19-69 80TH STREET 11370  
## RESIDENTIAL.UNITS COMMERCIAL.UNITS TOTAL.UNITS LAND.SQUARE.FEET  
## 1 1 - 1 1,800   
## 2 1 - 1 1,800   
## 3 1 - 1 1,800   
## 4 1 - 1 1,800   
## 5 2 - 2 1,800   
## 6 3 - 3 2,000   
## GROSS.SQUARE.FEET YEAR.BUILT TAX.CLASS.AT.TIME.OF.SALE  
## 1 1,224 1950 1  
## 2 1,224 1950 1  
## 3 1,224 1950 1  
## 4 1,224 1950 1  
## 5 1,224 1950 1  
## 6 2,835 1945 1  
## BUILDING.CLASS.AT.TIME.OF.SALE SALE.PRICE SALE.DATE  
## 1 A5 600,000 4/10/2017  
## 2 A5 600,000 4/10/2017  
## 3 A5 660,000 7/26/2016  
## 4 A5 275,500 11/18/2016  
## 5 B1 - 6/13/2016  
## 6 C0 - 8/15/2016

tail(queens)

## BOROUGH NEIGHBORHOOD  
## 25822 4 WOODSIDE   
## 25823 4 WOODSIDE   
## 25824 4 WOODSIDE   
## 25825 4 WOODSIDE   
## 25826 4 WOODSIDE   
## 25827 4 WOODSIDE   
## BUILDING.CLASS.CATEGORY TAX.CLASS.AT.PRESENT  
## 25822 44 CONDO PARKING 4  
## 25823 44 CONDO PARKING 4  
## 25824 44 CONDO PARKING 4  
## 25825 44 CONDO PARKING 4  
## 25826 46 CONDO STORE BUILDINGS 4  
## 25827 46 CONDO STORE BUILDINGS 4  
## BLOCK LOT EASE.MENT BUILDING.CLASS.AT.PRESENT  
## 25822 2324 1160 NA RG  
## 25823 2324 1161 NA RG  
## 25824 2324 1162 NA RG  
## 25825 2324 1163 NA RG  
## 25826 2324 1007 NA RK  
## 25827 2324 1008 NA RK  
## ADDRESS APARTMENT.NUMBER ZIP.CODE  
## 25822 63-14 QUEENS BOULEVARD P-57 11377  
## 25823 63-14 QUEENS BOULEVARD P-58 11377  
## 25824 63-14 QUEENS BOULEVARD P-59 11377  
## 25825 63-14 QUEENS BOULEVARD P-60 11377  
## 25826 63-14 QUEENS BOULEVARD COMM7 11377  
## 25827 63-14 QUEENS BOULEVARD COMM8 11377  
## RESIDENTIAL.UNITS COMMERCIAL.UNITS TOTAL.UNITS LAND.SQUARE.FEET  
## 25822 0 0 1 -   
## 25823 0 0 1 -   
## 25824 0 0 1 -   
## 25825 0 0 1 -   
## 25826 0 0 1 -   
## 25827 0 0 1 -   
## GROSS.SQUARE.FEET YEAR.BUILT TAX.CLASS.AT.TIME.OF.SALE  
## 25822 - 2008 4  
## 25823 - 2008 4  
## 25824 - 2008 4  
## 25825 - 2008 4  
## 25826 - 2008 4  
## 25827 - 2008 4  
## BUILDING.CLASS.AT.TIME.OF.SALE SALE.PRICE SALE.DATE  
## 25822 RG - 3/17/2017  
## 25823 RG - 3/17/2017  
## 25824 RG - 3/17/2017  
## 25825 RG - 3/17/2017  
## 25826 RK 831,250 3/15/2017  
## 25827 RK - 3/15/2017

str(queens)

## 'data.frame': 25827 obs. of 21 variables:  
## $ BOROUGH : int 4 4 4 4 4 4 4 4 4 4 ...  
## $ NEIGHBORHOOD : Factor w/ 59 levels "AIRPORT LA GUARDIA ",..: 1 1 1 1 1 1 1 1 2 2 ...  
## $ BUILDING.CLASS.CATEGORY : Factor w/ 44 levels "01 ONE FAMILY DWELLINGS ",..: 1 1 1 1 2 3 12 12 1 1 ...  
## $ TAX.CLASS.AT.PRESENT : Factor w/ 11 levels " ","1","1A",..: 2 2 2 2 2 2 6 6 2 2 ...  
## $ BLOCK : int 976 976 976 976 976 949 949 949 15829 15830 ...  
## $ LOT : int 10 11 61 63 70 56 1012 1025 22 20 ...  
## $ EASE.MENT : logi NA NA NA NA NA NA ...  
## $ BUILDING.CLASS.AT.PRESENT : Factor w/ 127 levels " ","A0","A1",..: 7 7 7 7 11 15 91 91 3 3 ...  
## $ ADDRESS : Factor w/ 22698 levels "-00 136TH AVENUE ",..: 9071 9081 9121 9102 9051 8315 8307 19194 12682 12432 ...  
## $ APARTMENT.NUMBER : Factor w/ 1161 levels " ",..: 1 1 1 1 1 1 254 2 1 1 ...  
## $ ZIP.CODE : int 11370 11370 11370 11370 11370 11370 11370 11370 11691 11691 ...  
## $ RESIDENTIAL.UNITS : Factor w/ 88 levels " - "," 1 ",..: 2 2 2 2 3 4 2 2 2 2 ...  
## $ COMMERCIAL.UNITS : Factor w/ 27 levels " - "," 1 ",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ TOTAL.UNITS : Factor w/ 99 levels " - "," 1 ",..: 2 2 2 2 3 4 2 2 2 2 ...  
## $ LAND.SQUARE.FEET : Factor w/ 3236 levels " - "," 1,000 ",..: 369 369 369 369 369 722 1 1 1213 1071 ...  
## $ GROSS.SQUARE.FEET : Factor w/ 2917 levels " - "," 1,000 ",..: 196 196 196 196 196 1661 1 1 440 436 ...  
## $ YEAR.BUILT : int 1950 1950 1950 1950 1950 1945 0 0 2005 2005 ...  
## $ TAX.CLASS.AT.TIME.OF.SALE : int 1 1 1 1 1 1 2 2 1 1 ...  
## $ BUILDING.CLASS.AT.TIME.OF.SALE: Factor w/ 125 levels "A0","A1","A2",..: 6 6 6 6 10 14 89 89 2 2 ...  
## $ SALE.PRICE : Factor w/ 3285 levels " - "," 1 ",..: 2397 2397 2564 1184 1 1 1830 1448 1373 1619 ...  
## $ SALE.DATE : Factor w/ 355 levels "1/1/2017","1/10/2017",..: 182 182 284 72 241 302 9 69 305 51 ...

summary(queens)

## BOROUGH NEIGHBORHOOD   
## Min. :4 FLUSHING-NORTH : 2712   
## 1st Qu.:4 ASTORIA : 1181   
## Median :4 BAYSIDE : 1131   
## Mean :4 FOREST HILLS : 1052   
## 3rd Qu.:4 JACKSON HEIGHTS : 947   
## Max. :4 FLUSHING-SOUTH : 894   
## (Other) :17910   
## BUILDING.CLASS.CATEGORY  
## 01 ONE FAMILY DWELLINGS :8246   
## 02 TWO FAMILY DWELLINGS :5588   
## 10 COOPS - ELEVATOR APARTMENTS :3797   
## 13 CONDOS - ELEVATOR APARTMENTS :1728   
## 09 COOPS - WALKUP APARTMENTS :1250   
## 03 THREE FAMILY DWELLINGS :1249   
## (Other) :3969   
## TAX.CLASS.AT.PRESENT BLOCK LOT EASE.MENT   
## 1 :15160 Min. : 13 Min. : 1 Mode:logical   
## 2 : 7292 1st Qu.: 2783 1st Qu.: 16 NA's:25827   
## 4 : 1727 Median : 5893 Median : 39   
## 2A : 577 Mean : 6565 Mean : 209   
## 1B : 381 3rd Qu.: 9946 3rd Qu.: 81   
## 1A : 366 Max. :16322 Max. :8009   
## (Other): 324   
## BUILDING.CLASS.AT.PRESENT  
## A1 : 3809   
## D4 : 3797   
## A5 : 1992   
## B3 : 1947   
## B2 : 1841   
## R4 : 1645   
## (Other):10796   
## ADDRESS APARTMENT.NUMBER  
## 131-05 40TH ROAD : 145 :22850   
## 136-21 LATIMER PLACE : 79 2A : 53   
## 31-35 31ST STREET : 61 2B : 51   
## 42-60 CRESCENT STREET : 56 3B : 48   
## 34-32 35TH STREET : 52 3A : 46   
## 14-34 110TH STREET : 45 4A : 40   
## (Other) :25389 (Other) : 2739   
## ZIP.CODE RESIDENTIAL.UNITS COMMERCIAL.UNITS TOTAL.UNITS   
## Min. : 0 1 :10485 0 :24001 1 :11527   
## 1st Qu.:11360 0 : 7437 1 : 977 0 : 6029   
## Median :11375 2 : 5324 - : 500 2 : 5252   
## Mean :11293 3 : 1289 2 : 159 3 : 1549   
## 3rd Qu.:11419 2 : 360 3 : 54 2 : 353   
## Max. :11697 4 : 279 4 : 32 4 : 314   
## (Other): 653 (Other): 104 (Other): 803   
## LAND.SQUARE.FEET GROSS.SQUARE.FEET YEAR.BUILT   
## - : 8524 - : 9167 Min. : 0   
## 4,000 : 1917 1,224 : 175 1st Qu.:1925   
## 2,500 : 1302 1,600 : 169 Median :1945   
## 2,000 : 1249 1,280 : 139 Mean :1846   
## 3,000 : 675 1,440 : 133 3rd Qu.:1960   
## 1,800 : 428 1,200 : 129 Max. :2016   
## (Other):11732 (Other):15915   
## TAX.CLASS.AT.TIME.OF.SALE BUILDING.CLASS.AT.TIME.OF.SALE  
## Min. :1.000 A1 : 3813   
## 1st Qu.:1.000 D4 : 3797   
## Median :1.000 A5 : 1993   
## Mean :1.521 B3 : 1930   
## 3rd Qu.:2.000 B2 : 1835   
## Max. :4.000 R4 : 1728   
## (Other):10731   
## SALE.PRICE SALE.DATE   
## - : 8035 11/10/2016: 182   
## 10 : 204 6/30/2016 : 174   
## 650,000 : 155 11/22/2016: 165   
## 450,000 : 151 10/28/2016: 163   
## 250,000 : 150 12/8/2016 : 162   
## 600,000 : 136 12/22/2016: 157   
## (Other) :16996 (Other) :24824

Clean the data

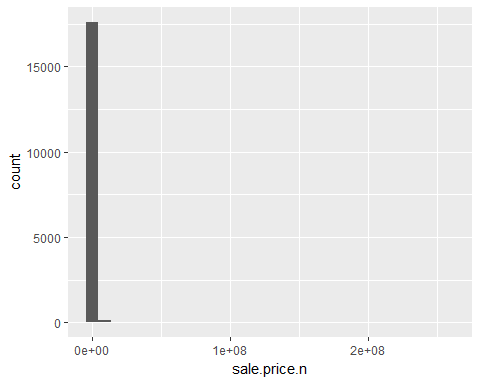
queens$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","", queens$SALE.PRICE))  
  
#Check the number of NAs that exist  
count(is.na(queens$SALE.PRICE.N))

## x freq  
## 1 FALSE 17792  
## 2 TRUE 8035

names(queens) <- tolower(names(queens)) # make all variable names lower case  
  
## Get rid of leading digits  
queens$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", queens$gross.square.feet))  
queens$land.sqft <- as.numeric(gsub("[^[:digit:]]","", queens$land.square.feet))  
queens$year.built <- as.numeric(as.character(queens$year.built))  
  
attach(queens)  
#hist(sale.price.n)   
qplot(sale.price.n, geom="histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

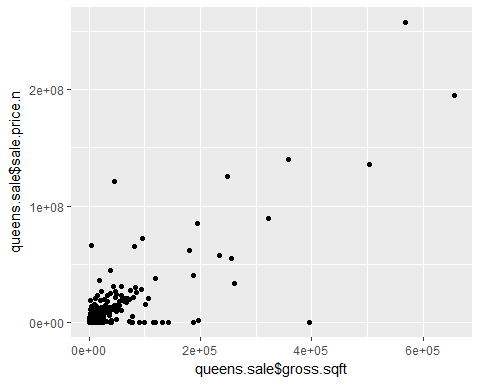
## Warning: Removed 8035 rows containing non-finite values (stat\_bin).



detach(queens)  
## keep only the actual sales  
  
queens.sale <- queens[queens$sale.price.n!=0,]

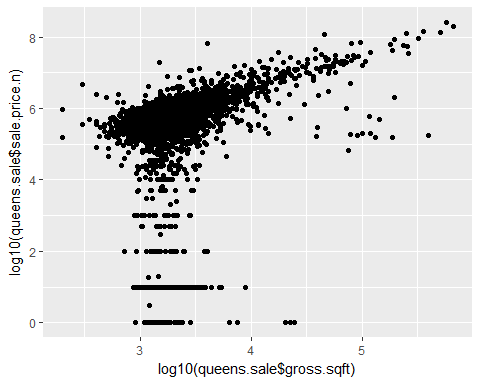
#plot(queens.sale$gross.sqft,queens.sale$sale.price.n)  
p <- ggplot(queens, aes(queens.sale$gross.sqft, queens.sale$sale.price.n))  
p + geom\_point()

## Warning: Removed 15082 rows containing missing values (geom\_point).



#plot log10(queens.sale$gross.sqft,queens.sale$sale.price.n)  
y <- ggplot(queens, aes(log10(queens.sale$gross.sqft), log10(queens.sale$sale.price.n)))  
y + geom\_point()

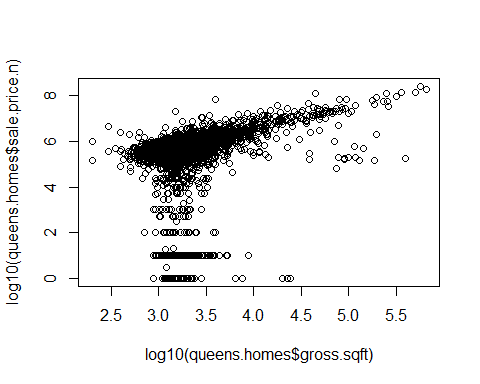
## Warning: Removed 15082 rows containing missing values (geom\_point).



## for now, let's look at 1-, 2-, and 3-family homes  
queens.homes <- queens.sale#[which(grepl("FAMILY",queens.sale$building.class.category)),]  
dim(queens.homes)

## [1] 25827 24

plot(log10(queens.homes$gross.sqft),log10(queens.homes$sale.price.n))



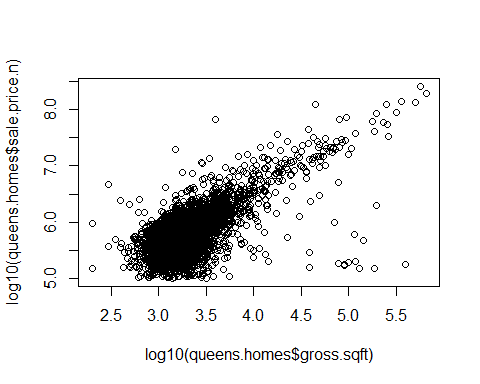
summary(queens.homes[which(queens.homes$sale.price.n<100000),])

## borough neighborhood  
## Min. :4 FLUSHING-NORTH : 60   
## 1st Qu.:4 FLUSHING-SOUTH : 37   
## Median :4 HOWARD BEACH : 35   
## Mean :4 SOUTH JAMAICA : 33   
## 3rd Qu.:4 JACKSON HEIGHTS : 30   
## Max. :4 ASTORIA : 27   
## (Other) :519   
## building.class.category  
## 01 ONE FAMILY DWELLINGS :224   
## 02 TWO FAMILY DWELLINGS :151   
## 10 COOPS - ELEVATOR APARTMENTS : 98   
## 44 CONDO PARKING : 84   
## 05 TAX CLASS 1 VACANT LAND : 48   
## 03 THREE FAMILY DWELLINGS : 33   
## (Other) :103   
## tax.class.at.present block lot ease.ment   
## 1 :409 Min. : 15 Min. : 1.0 Mode:logical   
## 2 :156 1st Qu.: 3524 1st Qu.: 18.0 NA's:741   
## 4 :109 Median : 9048 Median : 41.0   
## 1B : 48 Mean : 7973 Mean : 277.3   
## 2A : 8 3rd Qu.:11451 3rd Qu.: 120.0   
## 1A : 6 Max. :16229 Max. :8009.0   
## (Other): 5   
## building.class.at.present address   
## A1 :113 138-35 ELDER AVENUE : 8   
## D4 : 98 63-34 FRESH POND ROAD : 7   
## B3 : 72 11-24 31ST AVENUE : 5   
## RG : 59 10-11 NAMEOKE STREET : 4   
## A2 : 49 116-29 221ST STREET : 4   
## V0 : 47 30-80 42ND STREET : 4   
## (Other):303 (Other) :709   
## apartment.number zip.code residential.units commercial.units  
## :618 Min. : 0 0 :277 0 :698   
## P2 : 5 1st Qu.:11364 1 :259 - : 24   
## 2B : 4 Median :11411 2 :136 1 : 10   
## P12 : 4 Mean :10902 3 : 34 2 : 4   
## P4 : 4 3rd Qu.:11427 2 : 16 2 : 2   
## P5 : 4 Max. :11694 - : 8 1 : 1   
## (Other) :102 (Other): 11 (Other): 2   
## total.units land.square.feet gross.square.feet year.built   
## 1 :350 - :256 - :312 Min. : 0   
## 0 :180 4,000 : 56 1,200 : 7 1st Qu.:1925   
## 2 :139 2,000 : 48 1,120 : 6 Median :1940   
## 3 : 34 2,500 : 32 1,848 : 6 Mean :1714   
## 2 : 18 3,000 : 19 1,304 : 5 3rd Qu.:1960   
## - : 5 2,200 : 10 1,496 : 5 Max. :2016   
## (Other): 15 (Other):320 (Other):400   
## tax.class.at.time.of.sale building.class.at.time.of.sale sale.price   
## Min. :1.000 A1 :113 10 :204   
## 1st Qu.:1.000 D4 : 98 1 : 46   
## Median :1.000 B3 : 72 100 : 30   
## Mean :1.677 RG : 62 1,000 : 29   
## 3rd Qu.:2.000 A2 : 49 10,000 : 28   
## Max. :4.000 V0 : 47 25,000 : 25   
## (Other):300 (Other) :379   
## sale.date sale.price.n gross.sqft land.sqft   
## 9/15/2016 : 10 Min. : 1 Min. : 400 Min. : 300   
## 11/28/2016: 7 1st Qu.: 10 1st Qu.: 1296 1st Qu.: 2042   
## 12/12/2016: 7 Median :10000 Median : 1612 Median : 2850   
## 3/2/2017 : 7 Mean :26216 Mean : 2196 Mean : 3824   
## 3/27/2017 : 7 3rd Qu.:50000 3rd Qu.: 2128 3rd Qu.: 4000   
## 7/13/2016 : 7 Max. :98653 Max. :75368 Max. :54000   
## (Other) :696 NA's :312 NA's :256

""

## [1] ""

## remove outliers that seem like they weren't actual sales  
queens.homes$outliers <- (log10(queens.homes$sale.price.n) <=5) + 0  
queens.homes <- queens.homes[which(queens.homes$outliers==0),]  
plot(log10(queens.homes$gross.sqft),log10(queens.homes$sale.price.n))



plot((queens.homes$gross.sqft),(queens.homes$sale.price.n))

