**Manhattan\_sales\_1.Rproj Data Cleaning and exploration**

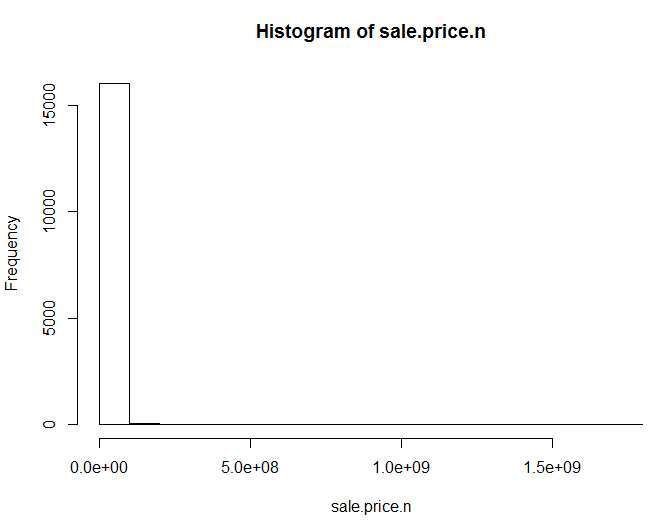
**2/26/2017, Robert Flamenbaum, Patrick McDevitt**

## keep only the actual sales

mn.sale <- mn[mn$sale.price.n!=0,]

plot(mn.sale$gross.sqft,mn.sale$sale.price.n)

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



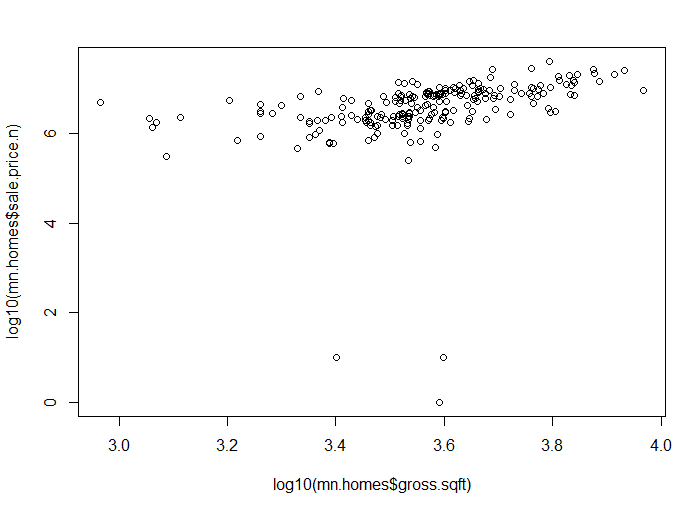
## for now, let's look at 1-, 2-, and 3-family homes

mn.homes <- mn.sale[which(grepl("FAMILY",mn.sale$building.class.category)),]

dim(mn.homes)

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

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

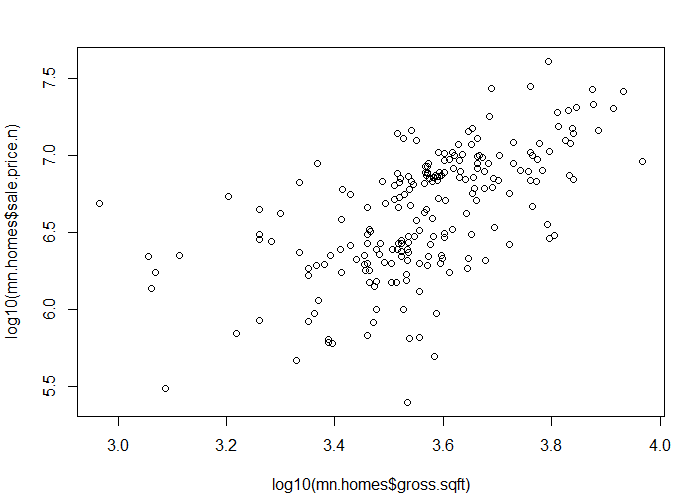


## remove outliers that seem like they weren't actual sales

mn.homes$outliers <- (log10(mn.homes$sale.price.n) <=5) + 0

mn.homes <- mn.homes[which(mn.homes$outliers==0),]

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



boxplot(log10(mn.homes$sale.price.n), data=mn, main="Manhattan Home Sale Price",

xlab="Mahattan Sales", ylab="Log Sale Price")

