Jovani Ntabgoba

May 9, 2016

# What happened after the 2008 US housing bubble? (a casestudy of Manhattan in 2009).

Stories about the 2008 Financial Crisis often mention which banks and companies that became bankrupt and people who lost jobs and homes. However, little is mentioned of what profits were made after,just after. This analysis explores a data set that contains 18673 houses sold in Manhattan between January 2009 to December 2009. The dataset is available at [NYC Website](www1.nyc.gov/assets/finance/downloads/.../2009_manhattan.xls)

### Loading required packages/libraries

# Set working dir and input data  
setwd("/Users/ntabgoba/Desktop/manhatta")  
library(readr)  
library(ggplot2)  
library(dplyr)  
library(caret)  
library(doMC)  
library(knitr)  
mhattan <- read.csv("2009\_manhattan.csv",header = TRUE, skip = 4)  
dim(mhattan)

## [1] 18673 20

head(mhattan,2)

## BOROUGH NEIGHBORHOOD  
## 1 1 ALPHABET CITY   
## 2 1 ALPHABET CITY   
## BUILDING.CLASS.CATEGORY TAX.CLASS.AT.PRESENT BLOCK  
## 1 07 RENTALS - WALKUP APARTMENTS 2B 372  
## 2 07 RENTALS - WALKUP APARTMENTS 2A 377  
## LOT EASE.MENT BUILDING.CLASS.AT.PRESENT  
## 1 19 NA C7  
## 2 53 NA C2  
## ADDRESS ZIP.CODE RESIDENTIAL.UNITS  
## 1 292 EAST 3 STREET 10009 9  
## 2 269 EAST 7 STREET 10009 5  
## COMMERCIAL.UNITS TOTAL.UNITS LAND.SQUARE.FEET GROSS.SQUARE.FEET  
## 1 1 10 2,401 6,920  
## 2 0 5 2,169 3,657  
## YEAR.BUILT TAX.CLASS.AT.TIME.OF.SALE BUILDING.CLASS.AT.TIME.OF.SALE  
## 1 1920 2 C7   
## 2 1900 2 C2   
## SALE.PRICE SALE.DATE  
## 1 $670,509 5/11/09  
## 2 $0 3/25/09

mhattan <- as\_data\_frame(mhattan)

### Data clearning and wrangling/munging

# Select columns relevant to analysis  
mhatta <- select(mhattan, NEIGHBORHOOD, TAX.CLASS.AT.PRESENT, BLOCK, BUILDING.CLASS.AT.PRESENT, ZIP.CODE,  
 RESIDENTIAL.UNITS, COMMERCIAL.UNITS, TOTAL.UNITS, LAND.SQUARE.FEET, GROSS.SQUARE.FEET,   
 YEAR.BUILT, TAX.CLASS.AT.TIME.OF.SALE, BUILDING.CLASS.AT.TIME.OF.SALE, SALE.PRICE, SALE.DATE)

# Rename the columns  
mhatt <- rename(mhatta, neighbd = NEIGHBORHOOD, tclass\_present = TAX.CLASS.AT.PRESENT, block = BLOCK ,   
 bclass\_present = BUILDING.CLASS.AT.PRESENT, zip = ZIP.CODE, res\_units = RESIDENTIAL.UNITS,  
 com\_units = COMMERCIAL.UNITS ,total\_units = TOTAL.UNITS, land\_ft = LAND.SQUARE.FEET,   
 gross\_ft = GROSS.SQUARE.FEET, year\_built = YEAR.BUILT, tclass\_sale = TAX.CLASS.AT.TIME.OF.SALE,   
 bclass\_sale = BUILDING.CLASS.AT.TIME.OF.SALE ,sale\_price = SALE.PRICE ,sale\_date = SALE.DATE)

\* After selecting the variables that are relevant to the analysis. We proceed to rename columns to shorter and easily programmeable wording.  
 \* We also change data types of price to numbers, tax to factors and date to date. R had previously automatically read price,tax and date as characters/strings.  
 \* We also add new variables that are a function of existing variables.

# change column data types to their actual types  
mhat <- transmute(mhatt,  
 neighbd,  
 tclass\_present,  
 block,  
 bclass\_present,  
 zip,  
 res\_units,  
 com\_units,  
 total\_units,  
 land\_ft = parse\_number(mhatt$land\_ft),  
 gross\_ft = parse\_number(mhatt$gross\_ft),  
 year\_built,  
 tclass\_sale = parse\_factor(mhatt$tclass\_sale,levels = c(1,2,3,4)),  
 bclass\_sale,  
 sale\_price = parse\_number(mhatt$sale\_price),  
 sale\_date = parse\_date(mhatt$sale\_date, "%m/%d/%y")  
)

# EXPLORATORY.

max(mhat$year\_built)

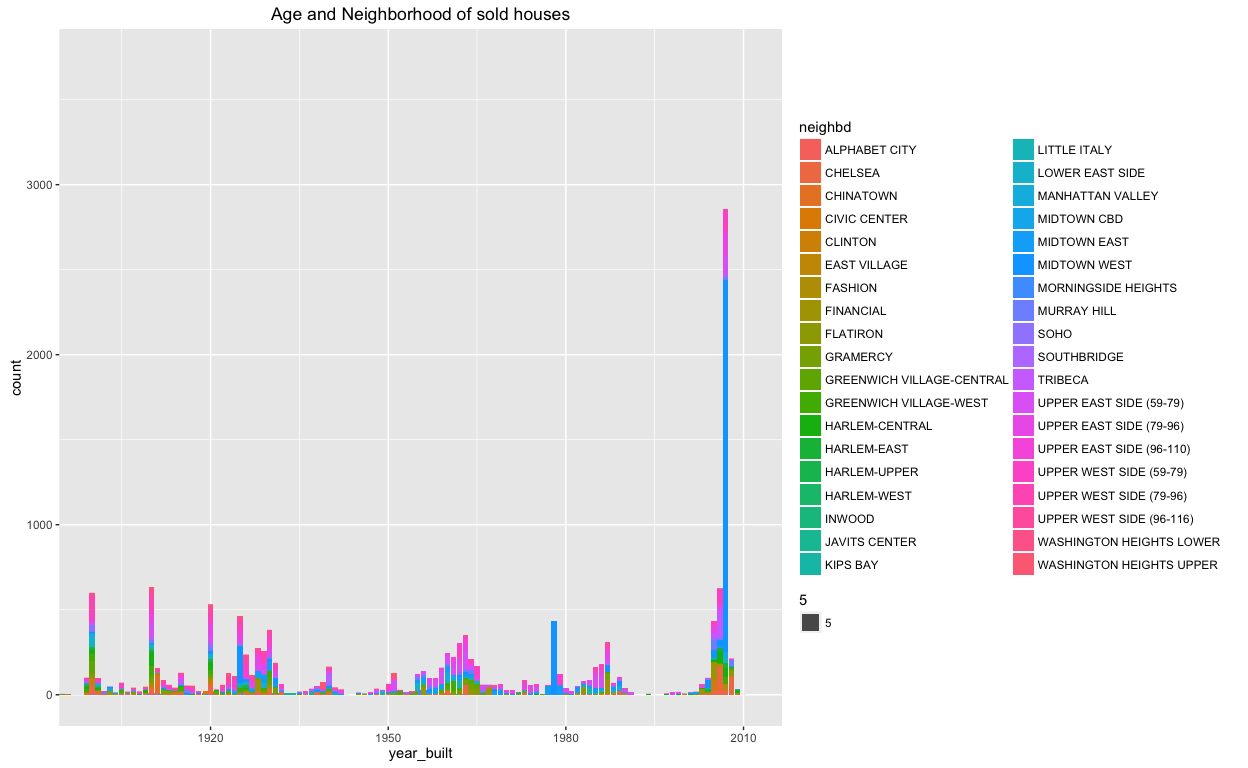
## [1] 2010

min(mhat$year\_built)

## [1] 0

### Age and neighborhood of most sold houses

ggplot(data = mhat) +  
 geom\_bar( mapping = aes(x = year\_built,fill = neighbd,size=5,na.rm = TRUE)) +  
 coord\_cartesian(xlim = c(1900,2011)) +  
 ggtitle("Age and Neighborhood of sold houses")

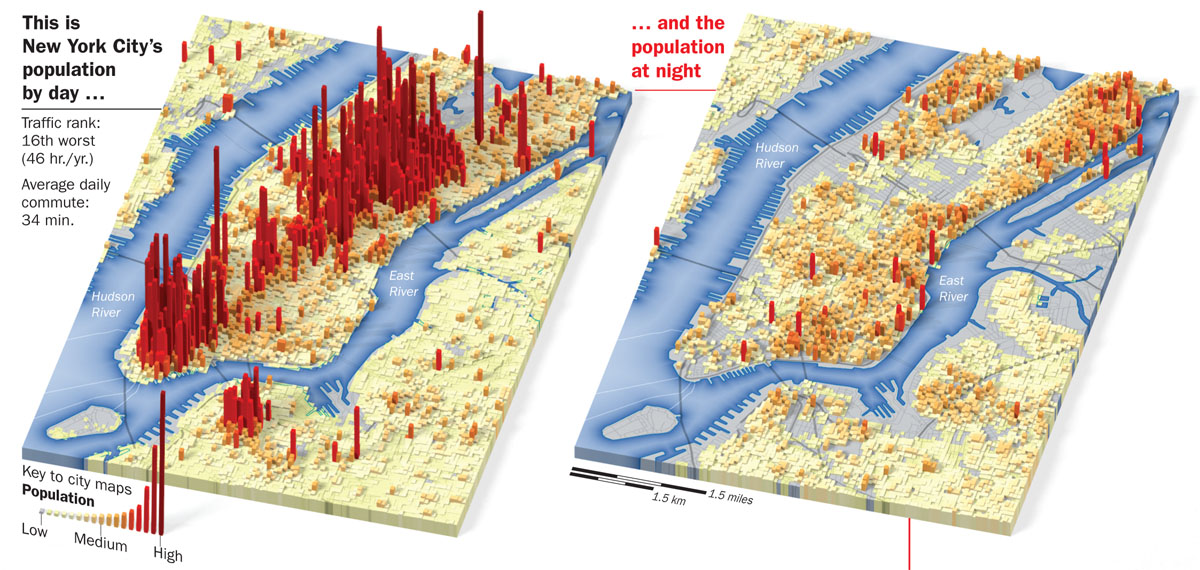
 Interestingly from the above graph 1, we see; - Houses built around 2007 were sold 3 times more than any other houses built in different periods. - Most of all other peaks of sales are also houses that were built during years that appear on the previous US Housing bubble timeline. Interesting this could draw us to conclude that the 2008 crisis, was also partly due to an aggregated aftershocks of speculated houses that were built due to incentives rather than pure demand and supply as depicted in below timeline.

knitr::include\_graphics("bubble.png")

### Commercial vs Residential

We investigate to see if as more residential(*expenditure*) houses increased, did commercial houses(*production*) increase in commensurate.

mha <- mhat %>%  
 filter(res\_units >= 1 & res\_units < 150,com\_units >=1)   
ggplot(data = mha, mapping = aes(x = res\_units, y = com\_units, color = tclass\_sale),na.rm = TRUE) +  
 geom\_point(position = "jitter") +  
 geom\_smooth(se = TRUE, color = "yellow") +  
 ggtitle("Commercial compared to Residential units, per Tax Class at Sale") +  
 facet\_wrap(~neighbd)

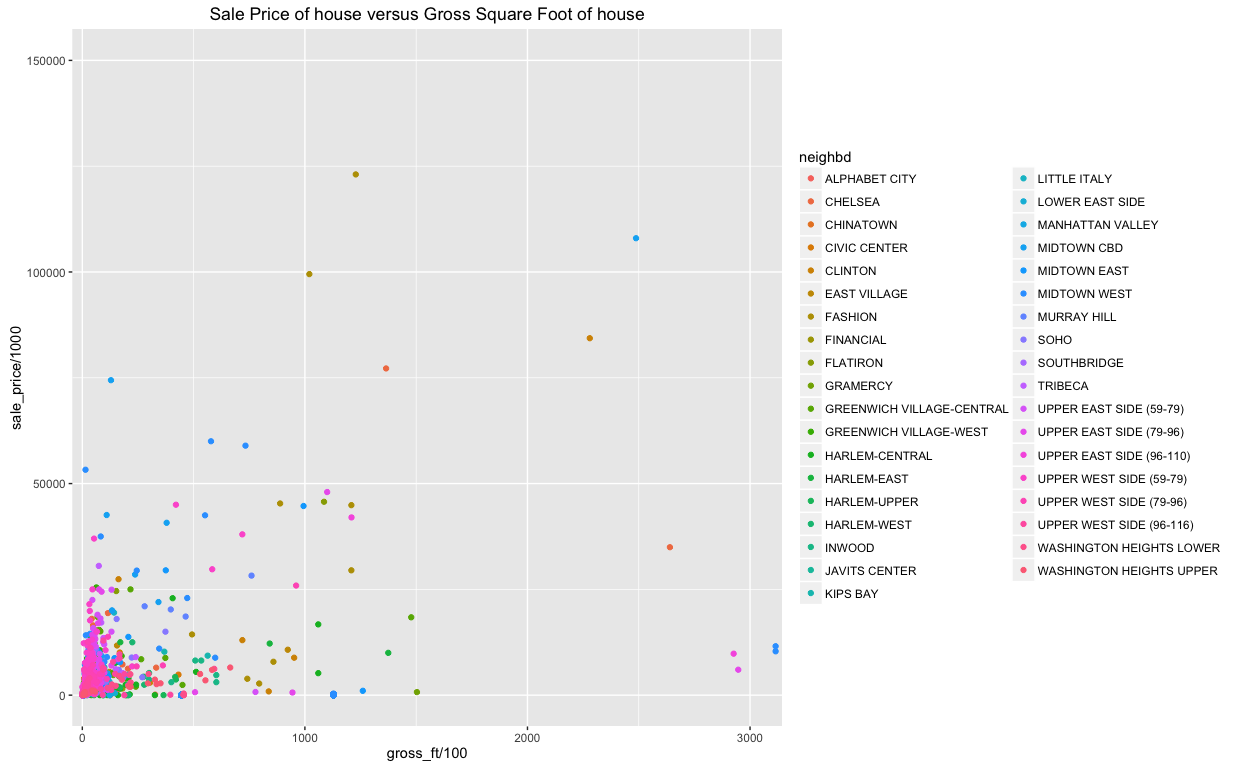
 From graph 2 above, we realise that; - More and more residential houses were built, while very few commercial houses were built. - Apartments and Condomiums that have more than 3 units were mainly built. Tax Class of sale 2 in green colour. - Upper East Side and West, Chelsea, East Village, Harlem-Central and Little Italy saw most of the sales. Interestingly the above layout has a relationship to this map of Manhattan's population during day & night.

### Relationship between price and gross area

# Filter out a $1.76Billion unit and freely offered apartments  
max(mhat$sale\_price)

## [1] 1.76e+09

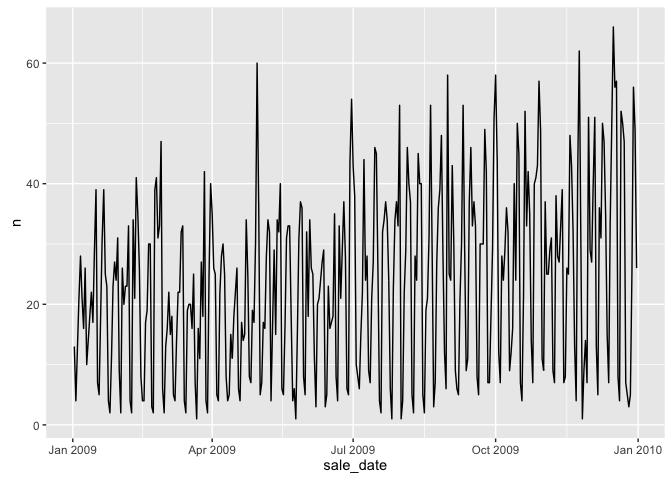
mha <- mhat %>%  
 filter(sale\_price >= 1000 & sale\_price < 1000000000, gross\_ft >=10)   
ggplot(data = mha) +  
 geom\_point(mapping = aes(x = gross\_ft/100,y = sale\_price/1000,color = neighbd, position = "jitter", na.rm = TRUE))+  
 ggtitle("Sale Price of house versus Gross Square Foot of house") +  
 coord\_cartesian(xlim = c(100,3000), ylim = c(100, 150000)) # Cut out over $100M units



Due to too much information in the plot, there is less we can discover from above graph 3.

# Relation of sales and period

monthly <- mha %>%  
 group\_by(sale\_date) %>%  
 summarise(n = n())  
ggplot(monthly, aes(sale\_date, n))+  
 geom\_line()

 From above graph 4, - We discover that volume of house sales, varies weekly due to weekends (may be). - We also realise that there was an upward increase in the volume of sales from Jan 2009 towards peak in Jan 2010. *NB:* Above graphs are simply exploratory analysis which helps us understand the data we are dealing with.

# MODEL APPLICATION

Model application is a process of trying to quantify the relationship between a variable of interest (predicted) versus other variable(s) known as predictors. ### Features to look at Sale\_Price, Gross\_ft, sale\_date, year\_built, neighbd, res\_units

mha\_model <- select(mha,neighbd,res\_units, year\_built, sale\_price, gross\_ft, sale\_date)%>%  
 filter(res\_units >= 1, gross\_ft >= 1)%>%  
 mutate(price\_per\_resUnit = sale\_price/res\_units,  
 sqFt\_per\_resUnit = gross\_ft/res\_units)

### Split into trainig and test data sets

class(mha\_model)

## [1] "tbl\_df" "tbl" "data.frame"

mha\_mod <- as.data.frame(mha\_model)  
class(mha\_mod)

## [1] "data.frame"

inTrain <- createDataPartition(y=mha\_mod$price\_per\_resUnit, p=0.7, list = FALSE)  
training <- mha\_mod[inTrain,]  
testing <- mha\_mod[-inTrain,]  
rbind("original data" = dim(mha\_mod), "training data" = dim(training), "testing data" = dim(testing))

## [,1] [,2]  
## original data 5440 8  
## training data 3809 8  
## testing data 1631 8

Above is the original data split into training and testing datasets.

# From exploratory analysis, year and total square foot of a unit are correlated to the price of a unit.  
cor(training$price\_per\_resUnit, training$sqFt\_per\_resUnit, use = "na")

## [1] 0.7094667

cor(training$price\_per\_resUnit,training$year\_built, use = "na")

## [1] 0.01514536

Price of a unit and Gross sqaure foot of a unit is highly correlated (0.7) and less correlated (0.0047) with the year which a unit was built.

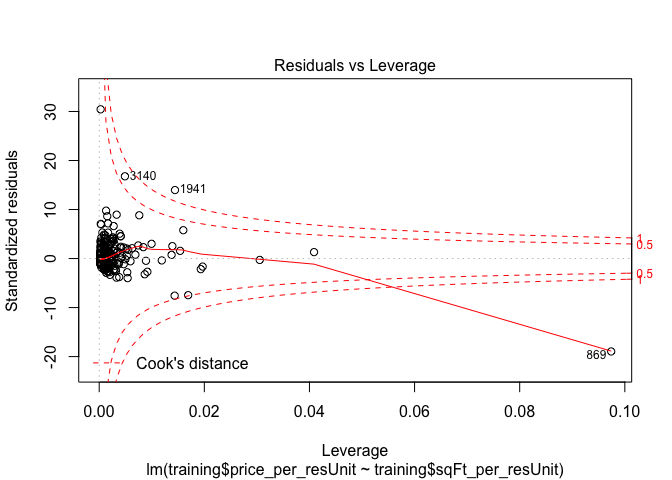
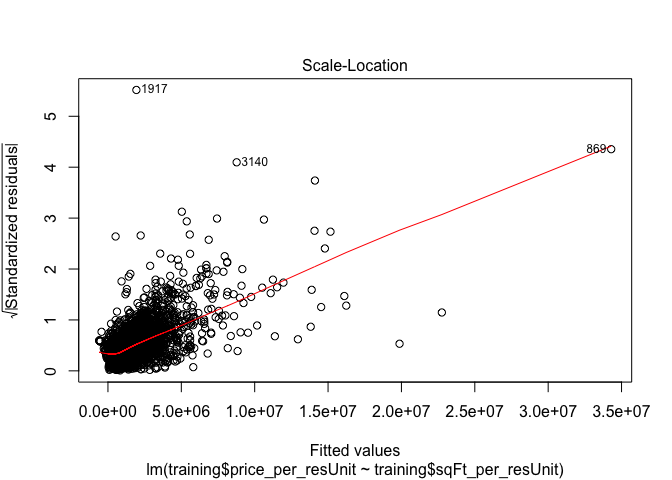
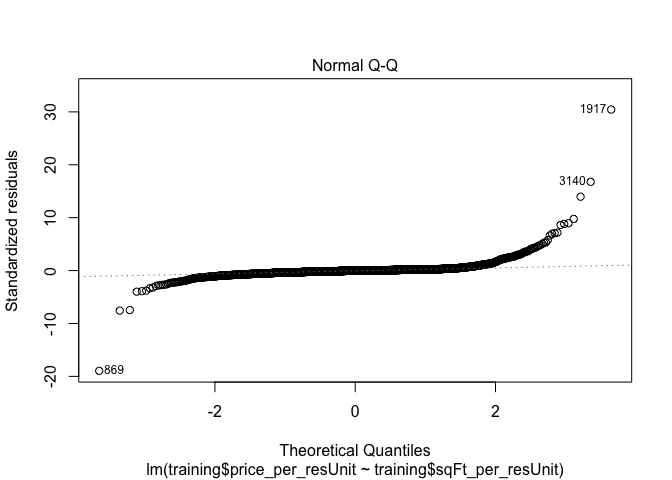
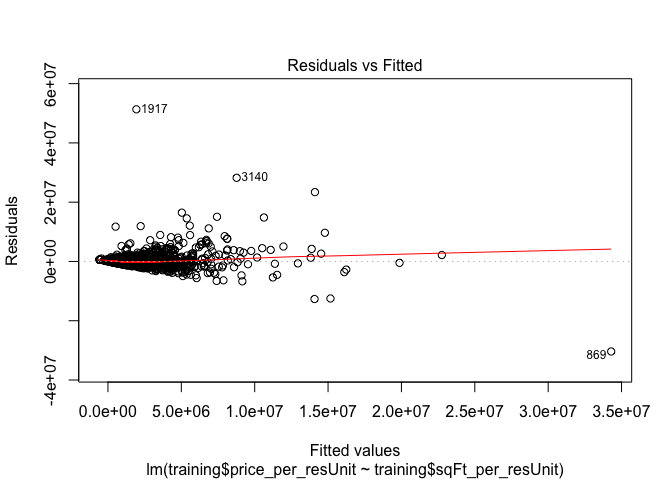
We use both year built and gross square foot to predict the price of unit house.

### Univariate linear regression

fit1 <- lm(training$price\_per\_resUnit ~ training$sqFt\_per\_resUnit)  
summary(fit1)

##   
## Call:  
## lm(formula = training$price\_per\_resUnit ~ training$sqFt\_per\_resUnit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30374971 -409987 -42853 221963 51331436   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -613954.78 45437.29 -13.51 <2e-16 \*\*\*  
## training$sqFt\_per\_resUnit 1778.82 28.64 62.12 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1687000 on 3807 degrees of freedom  
## Multiple R-squared: 0.5033, Adjusted R-squared: 0.5032   
## F-statistic: 3858 on 1 and 3807 DF, p-value: < 2.2e-16

plot(fit1)

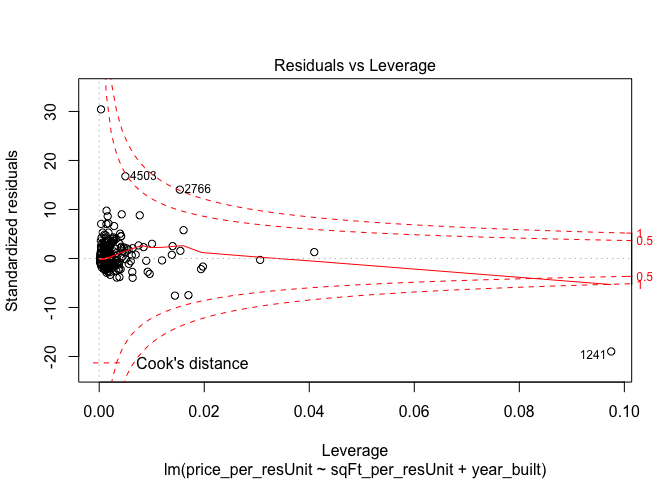
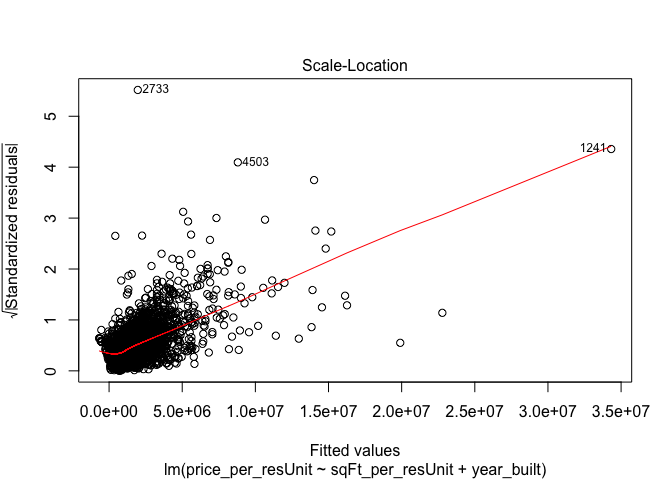
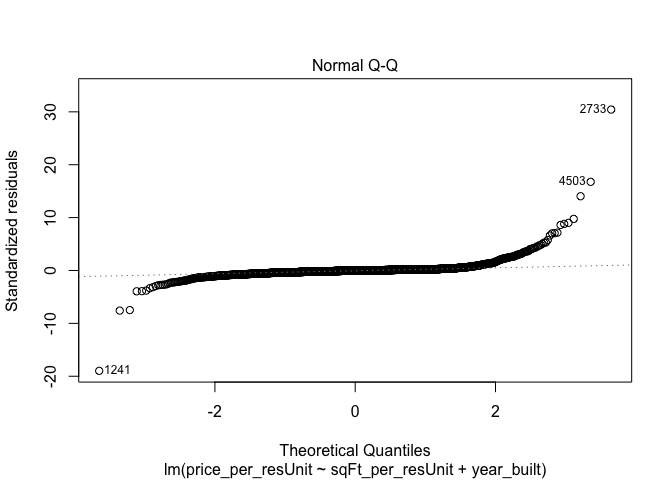
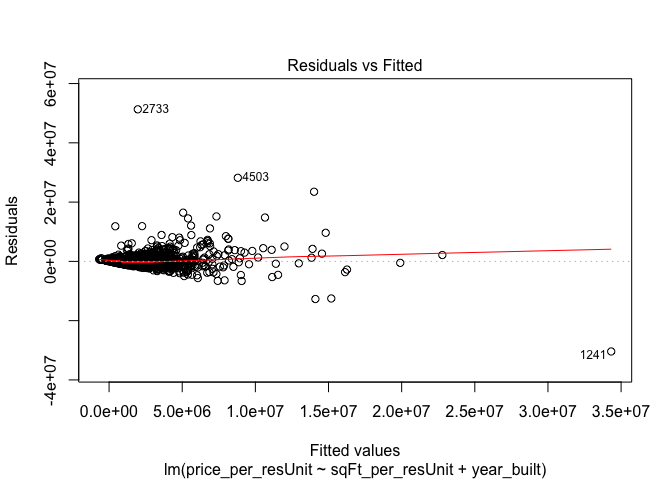


### Multivariate linear regression

fit11 <- lm(price\_per\_resUnit ~ sqFt\_per\_resUnit + year\_built, data = training)  
summary(fit11)

##   
## Call:  
## lm(formula = price\_per\_resUnit ~ sqFt\_per\_resUnit + year\_built,   
## data = training)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30403386 -413528 -47800 219468 51305564   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -710723.59 70130.77 -10.134 <2e-16 \*\*\*  
## sqFt\_per\_resUnit 1779.23 28.63 62.146 <2e-16 \*\*\*  
## year\_built 61.71 34.07 1.811 0.0702 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1686000 on 3806 degrees of freedom  
## Multiple R-squared: 0.5038, Adjusted R-squared: 0.5035   
## F-statistic: 1932 on 2 and 3806 DF, p-value: < 2.2e-16

plot(fit11)

 ### Random Forest

To be continued.....

END.