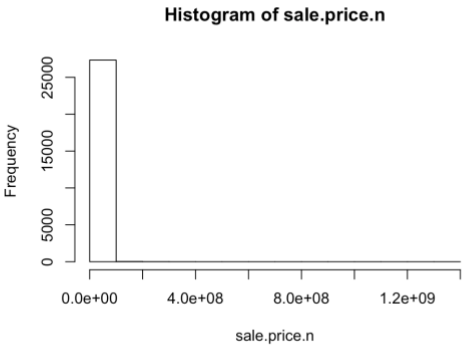
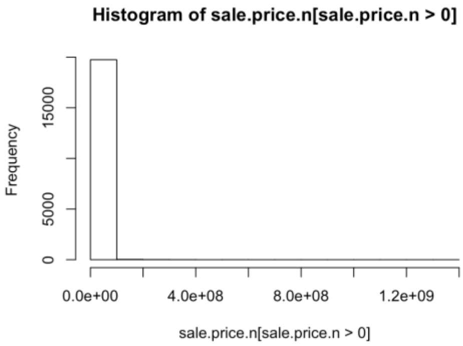
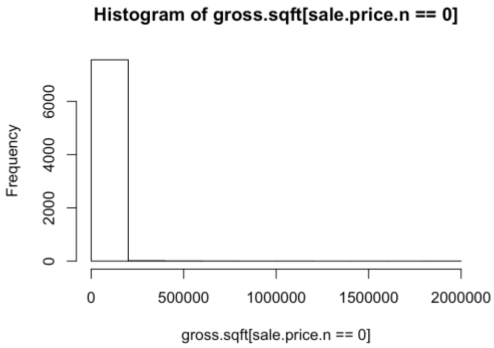
FE 582 Shuo Jin

Problem 1

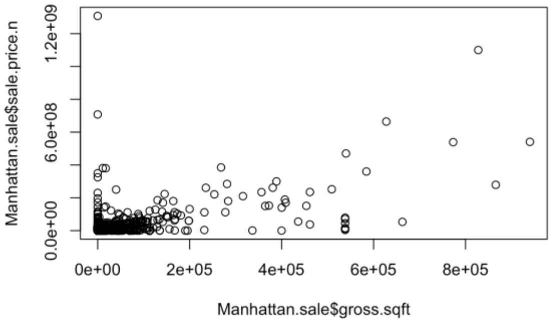
The Analysis of Manhattan

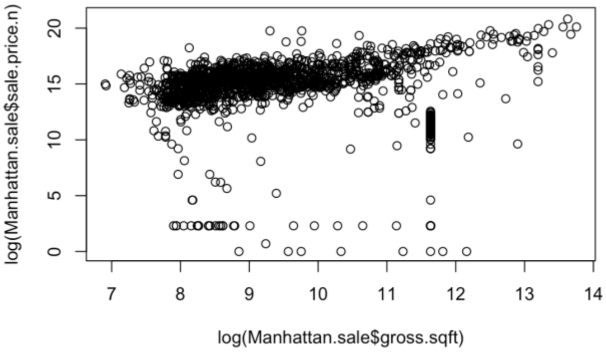




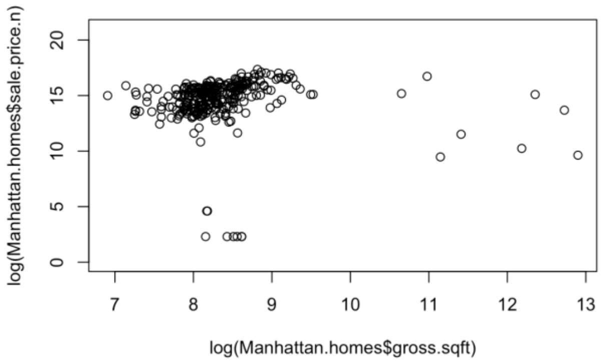


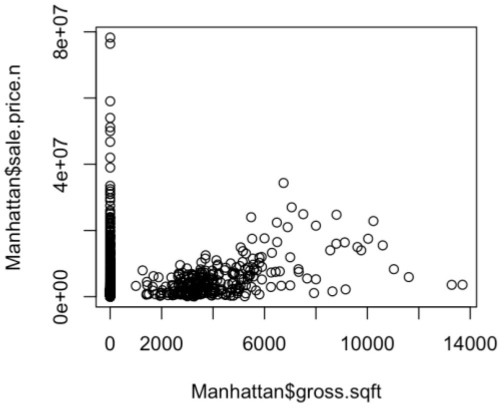
- *The Histogram of sale.price.n chart shows the data_Manhattan has 27395 rows*
- *The Histogram of sale.price.n[sale.price.n>0] shows the sale price large than 0 are 18802 rows which means there are 8593 price value are less than 0.*



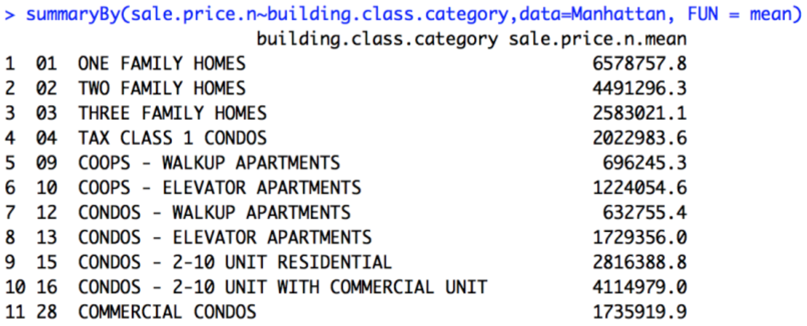


- *Only leave the actual sales only and show the plot of gross sqft and sale price. There still some points that the gross sqft is high but the price is low or the gross sqft is small but the sale price is high.*
- *There are several outliers which influence the distribution of the plot.*
- *Log them and see the relationship between these two variables.*

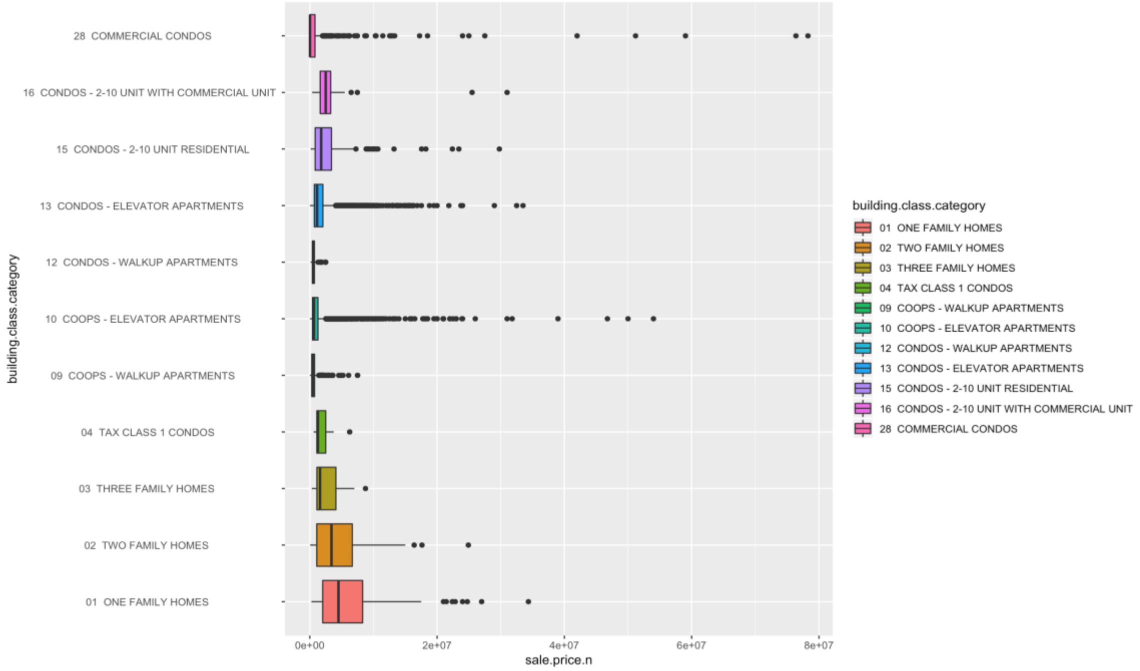




- *From above left chart we can make sure where is the outliers. We need to remove value which the log(Manhattan.homes$gross.sqft) >10 and the log(Manhattan.homes$sale.price.n) < 10, Which means the Manhattan.homes$gross.sqft > 100000 and the Manhattan.homes$sale.price.n < 100000.*
- *The above right chart is the Manhanttan data after clean up. But there must be a lot of missing gross sqft in the data set. The gross.sqft is 0 but have the sale price.*



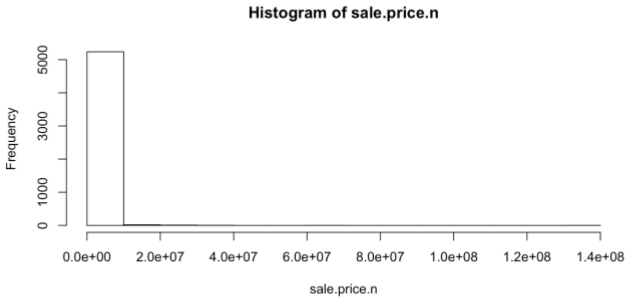
```
> summaryBy(sale.price.n~building.class.category,data=Manhattan, FUN = mean)
                        building.class.category sale.price.n.mean
1  01  ONE FAMILY HOMES                                  6578757.8
2  02  TWO FAMILY HOMES                                  4491296.3
3  03  THREE FAMILY HOMES                                2583021.1
4  04  TAX CLASS 1 CONDOS                                2022983.6
5  09  COOPS - WALKUP APARTMENTS                          696245.3
6  10  COOPS - ELEVATOR APARTMENTS                       1224054.6
7  12  CONDOS - WALKUP APARTMENTS                         632755.4
8  13  CONDOS - ELEVATOR APARTMENTS                      1729356.0
9  15  CONDOS - 2-10 UNIT RESIDENTIAL                    2816388.8
10 16  CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT           4114979.0
11 28  COMMERCIAL CONDOS                                 1735919.9
```

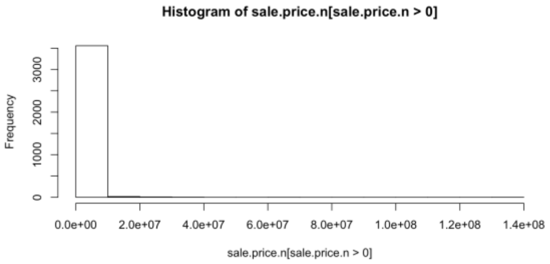


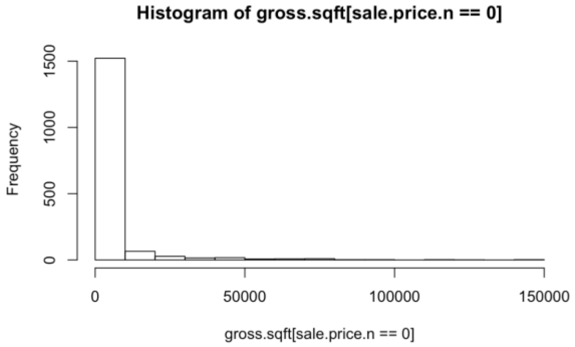
Here is the sale price of different building type boxplot in Manhattan

- *Overall the family mean price is larger than coops.*
- *The coops price is larger than Condos.*
- *Even though the **commercial condos** have the lowest mean price value.Tthrough the boxplot we can see that the price range of **commercial condos** is the largest*

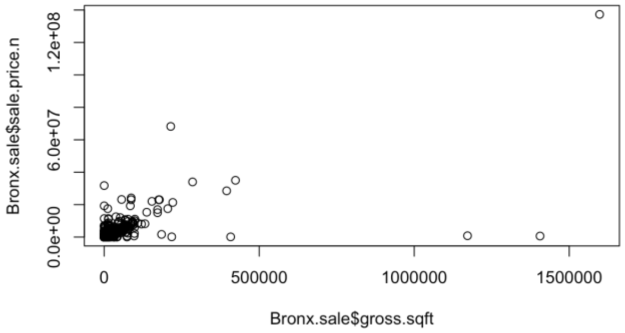
The Analysis of Bronx

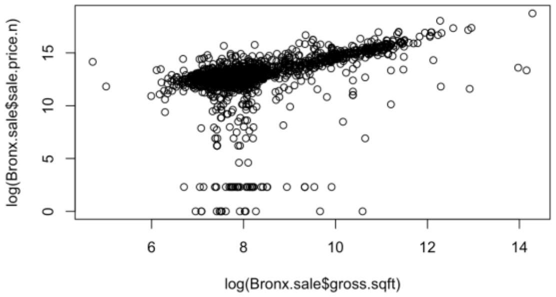




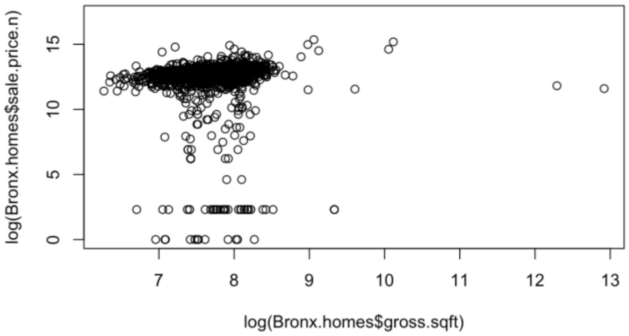


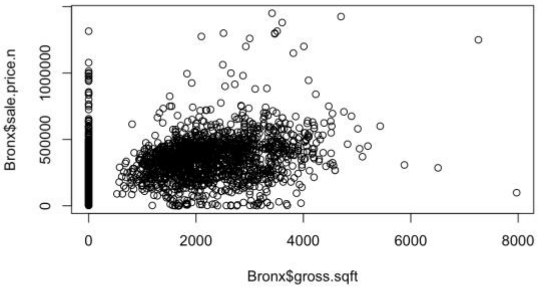
- *The Histogram of sale.price.n chart shows the data_Bronx has 5268 rows which is much smaller than Manhattan*
- *The Histogram of sale.price.n[sale.price.n>0] shows the sale price large than 0 are 4268 rows which means there are 1000 price value are less than 0.*



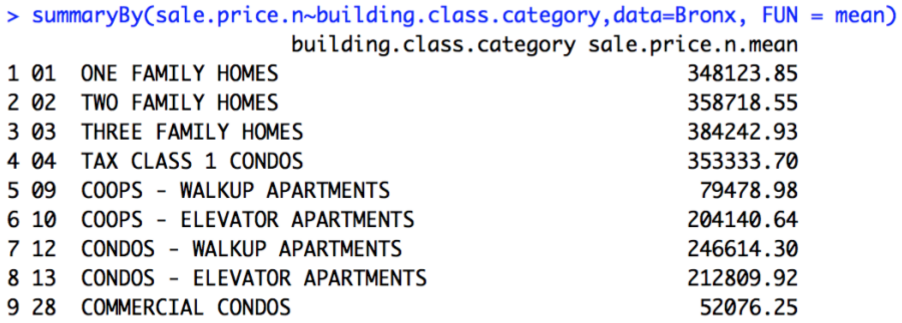


- *Only leave the actual sales only and show the plot of gross sqft and sale price. There still some points that the gross sqft is high but the price is low or the gross sqft is small but the sale price is high.*
- *There are several outliers which influence the distribution of the plot.*
- *Log them and see the relationship between these two variables.*

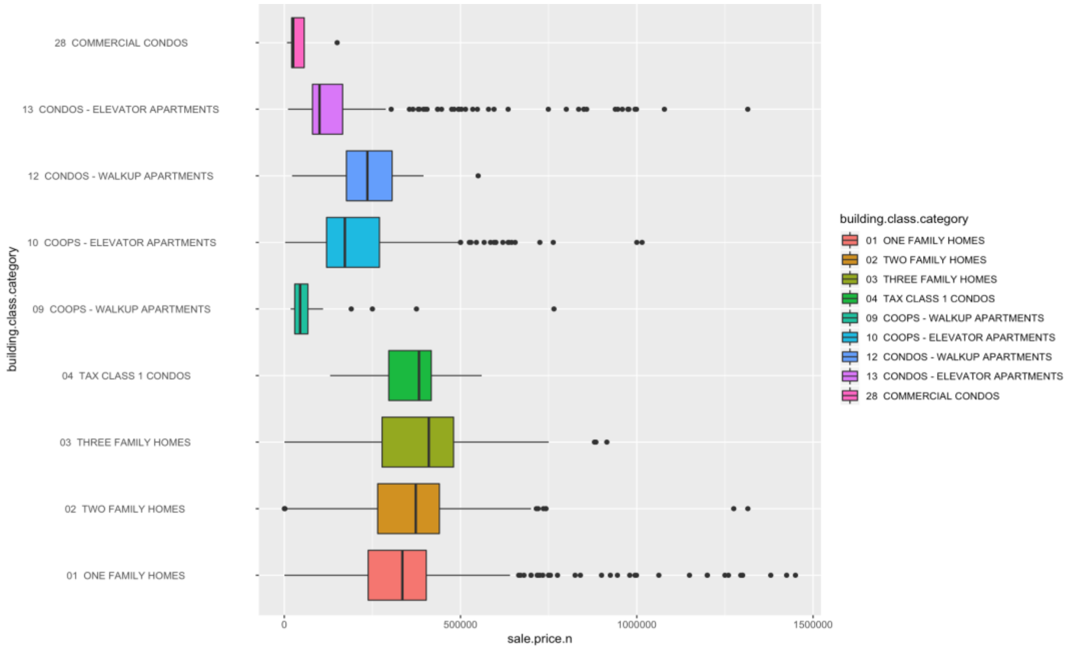




- *From above left chart we can make sure where is the outliers. We need to remove value which the log(Bronx.homes$gross.sqft) >10 and the log(Bronx.homes$sale.price.n) < 5, Which means the Bronx.homes$gross.sqft > 100000 and the Bronx.homes$sale.price.n < 100.*
- *The above right chart is the Bronx data after clean up. But there must be a lot of missing gross sqft in the data set. The gross.sqft is 0 but have the sale price.*



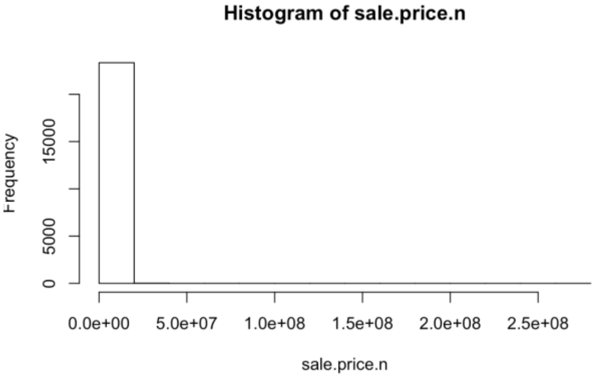
```
> summaryBy(sale.price.n~building.class.category,data=Bronx, FUN = mean)
                      building.class.category sale.price.n.mean
1 01  ONE FAMILY HOMES                                348123.85
2 02  TWO FAMILY HOMES                                358718.55
3 03  THREE FAMILY HOMES                              384242.93
4 04  TAX CLASS 1 CONDOS                              353333.70
5 09  COOPS - WALKUP APARTMENTS                        79478.98
6 10  COOPS - ELEVATOR APARTMENTS                     204140.64
7 12  CONDOS - WALKUP APARTMENTS                      246614.30
8 13  CONDOS - ELEVATOR APARTMENTS                    212809.92
9 28  COMMERCIAL CONDOS                                52076.25
```

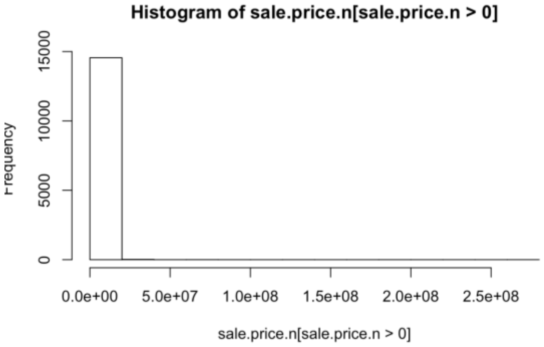


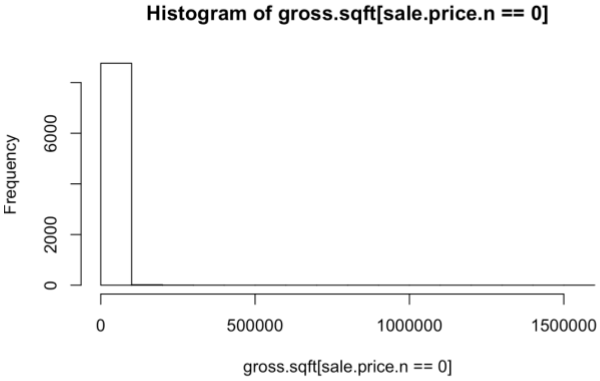
Here is the sale price of different building type boxplot in Bronx

- *Overall the family mean price is larger than coops.*
- *The coops price is larger than Condos.*
- *In Bronx the commercial condos does not have that much extremely high price.*
- *The one family price range is the largest.*

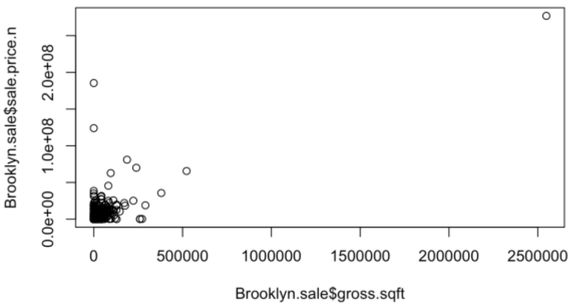
The Analysis of Brooklyn

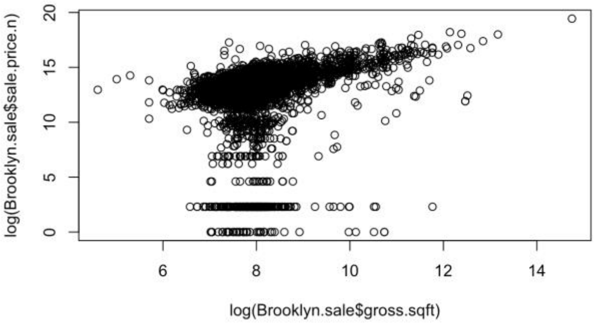




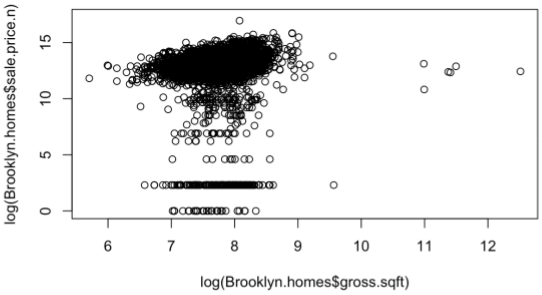


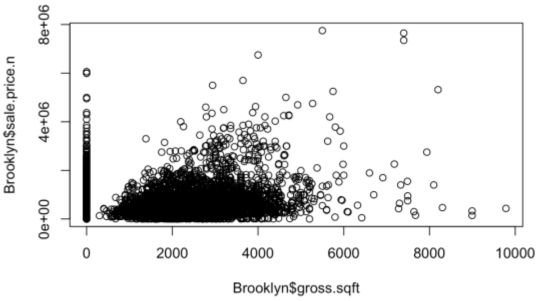
- *The Histogram of sale.price.n chart shows the data_Brooklyn has 23373 rows*
- *The Histogram of sale.price.n[sale.price.n>0] shows the sale price large than 0 are 13582 rows which means there are 9791 price value are less than 0.*



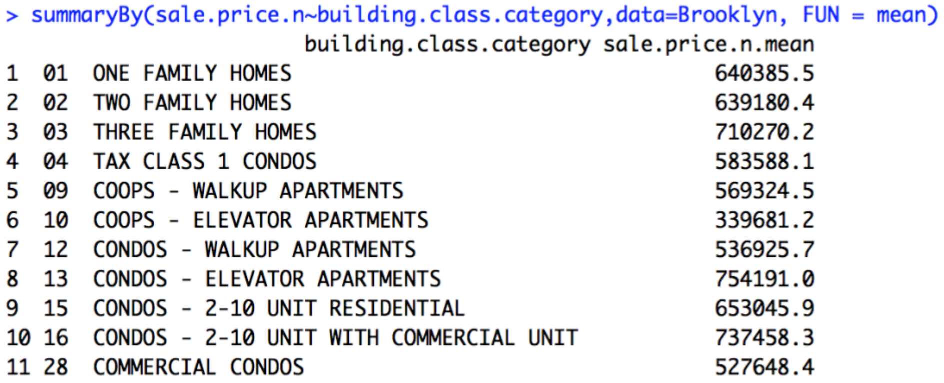


- *Only leave the actual sales only and show the plot of gross sqft and sale price. There still some points that the gross sqft is high but the price is low or the gross sqft is small but the sale price is high.*
- *There are several outliers which influence the distribution of the plot.*
- *Log them and see the relationship between these two variables.*

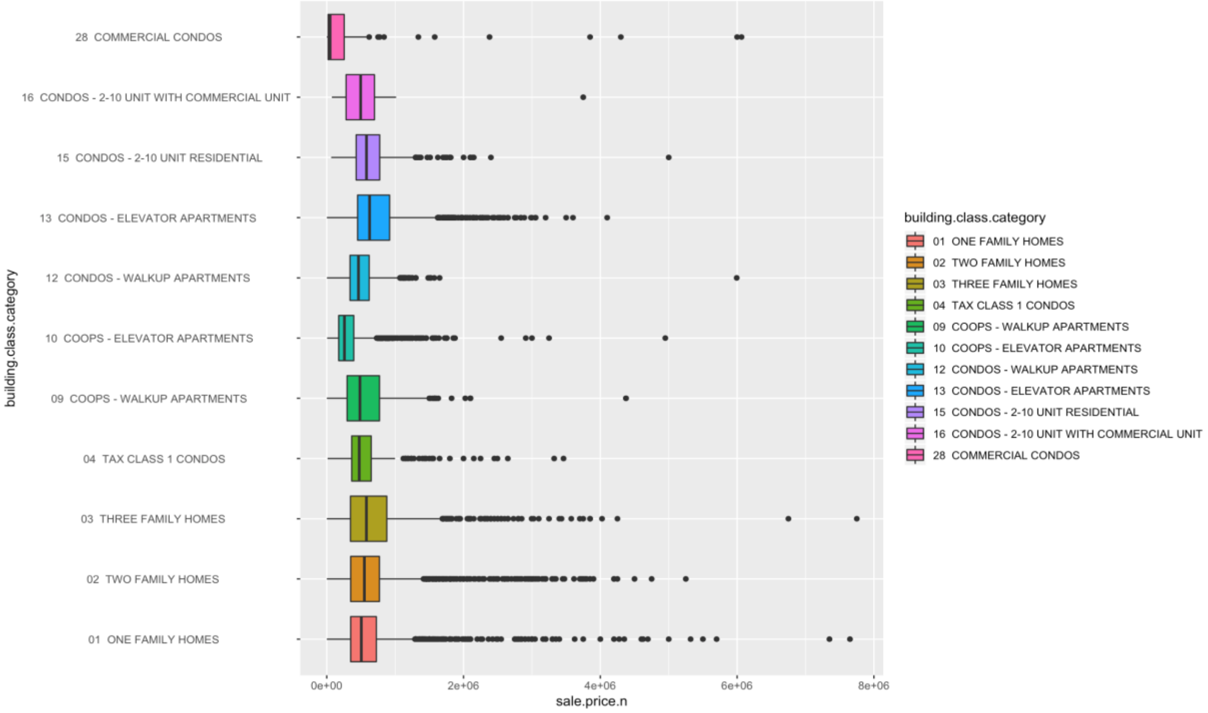




- *From above left chart we can make sure where is the outliers. We need to remove value which the log(Brooklyn.homes$gross.sqft) >10 and the log(Brooklyn.homes$sale.price.n) < 5, Which means the Brooklyn.homes$gross.sqft > 12000 and the Brooklyn.homes$sale.price.n < 100.*
- *The above right chart is the Bronx data after clean up. But there must be a lot of missing gross sqft in the data set. The gross.sqft is 0 but have the sale price.*



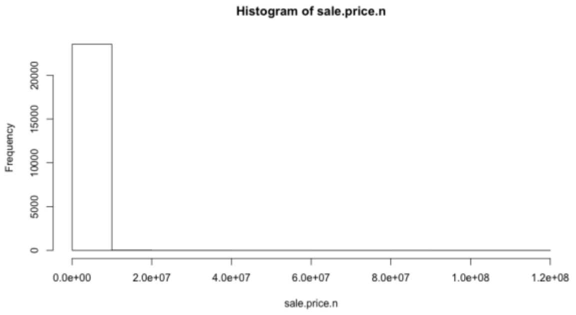
```
> summaryBy(sale.price.n~building.class.category,data=Brooklyn, FUN = mean)
                        building.class.category sale.price.n.mean
1  01  ONE FAMILY HOMES                                  640385.5
2  02  TWO FAMILY HOMES                                  639180.4
3  03  THREE FAMILY HOMES                                710270.2
4  04  TAX CLASS 1 CONDOS                                583588.1
5  09  COOPS - WALKUP APARTMENTS                         569324.5
6  10  COOPS - ELEVATOR APARTMENTS                       339681.2
7  12  CONDOS - WALKUP APARTMENTS                        536925.7
8  13  CONDOS - ELEVATOR APARTMENTS                      754191.0
9  15  CONDOS - 2-10 UNIT RESIDENTIAL                    653045.9
10 16  CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT           737458.3
11 28  COMMERCIAL CONDOS                                 527648.4
```

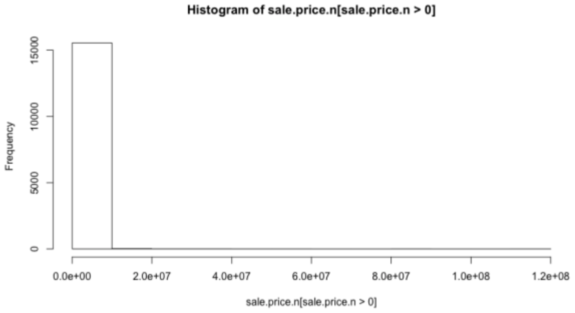


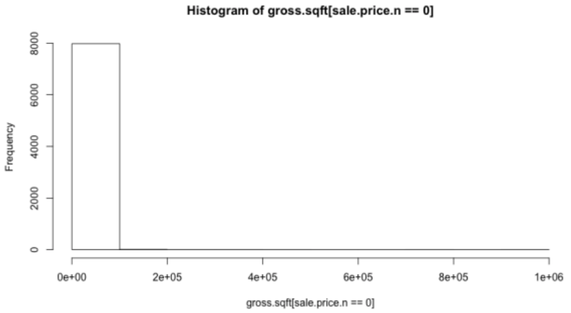
Here is the sale price of different building type boxplot in Brooklyn

- *In Brooklyn the family mean price is similar to Condos.*
- *In Brooklyn the commercial condos does not have that much extremely high price but also the lowest one.*
- *Every building type are have large price range means the price variance of Brooklyn is large.*

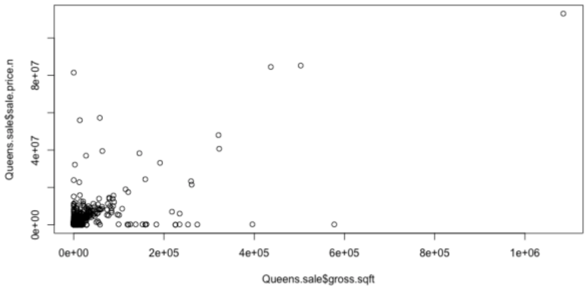
The Analysis of Queens

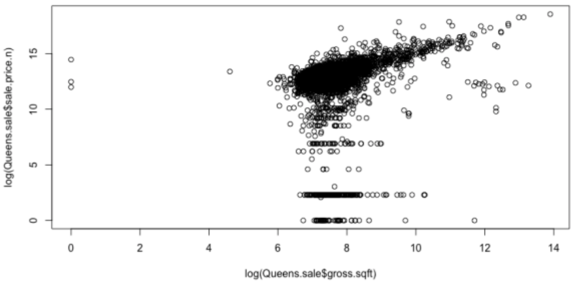




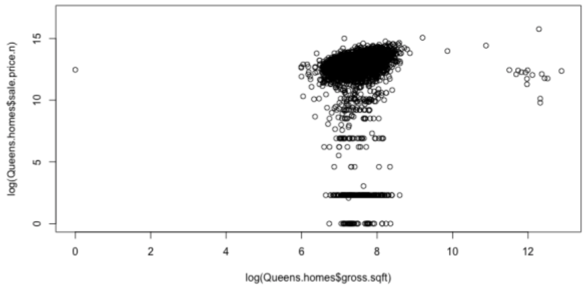


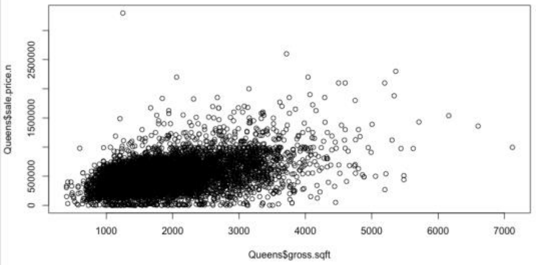
- *The Histogram of sale.price.n chart shows the data_Queens has 22583 rows*
- *The Histogram of sale.price.n[sale.price.n>0] shows the sale price large than 0 are 14587 rows which means there are 7996 price value are less than 0.*



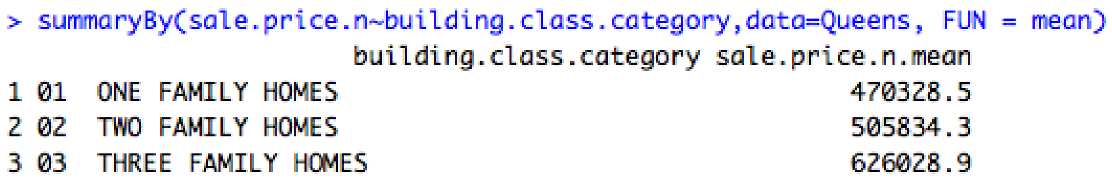


- *Only leave the actual sales only and show the plot of gross sqft and sale price. There still some points that the gross sqft is high but the price is low or the gross sqft is small but the sale price is high.*
- *There are several outliers which influence the distribution of the plot.*
- *Log them and see the relationship between these two variables.*

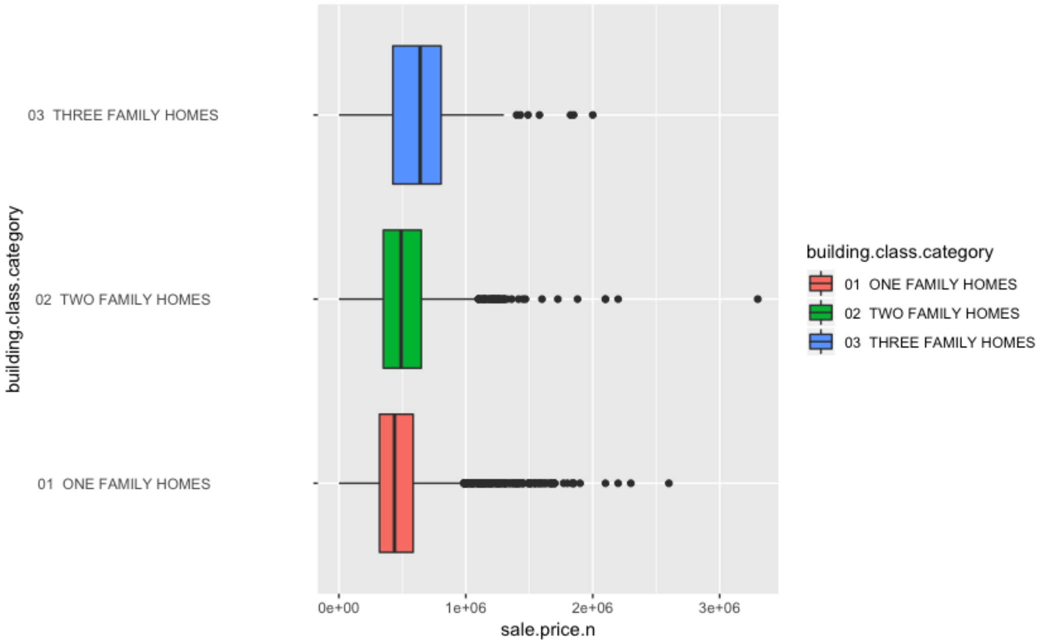




- *From above left chart we can make sure where is the outliers. We need to remove value which the log(Queens.homes$gross.sqft) >10 and the log(Queens.homes$sale.price.n) < 5, Which means the Queens.homes$gross.sqft > 8000 and the Queens.homes$sale.price.n < 100.*



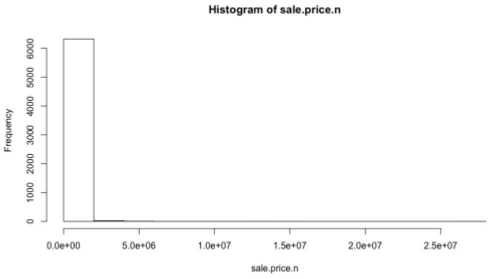
```
> summaryBy(sale.price.n~building.class.category,data=Queens, FUN = mean)
                     building.class.category sale.price.n.mean
1 01  ONE FAMILY HOMES                                470328.5
2 02  TWO FAMILY HOMES                                505834.3
3 03  THREE FAMILY HOMES                              626028.9
```

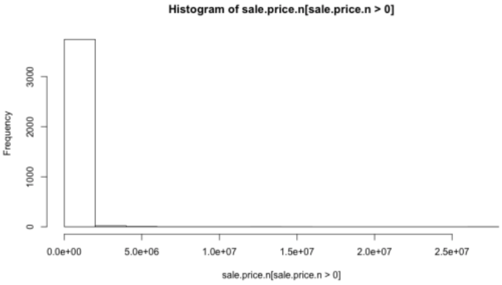


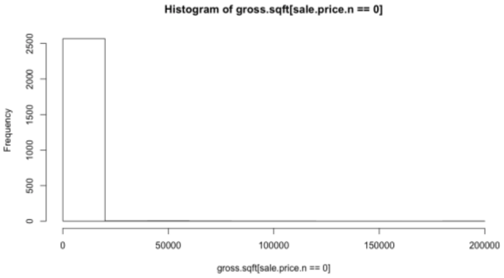
Here is the sale price of different building type boxplot in Queens

- *In Queens only have family type building*
- *The one family homes is the cheapest and the more the room the more expensive the family homes is*

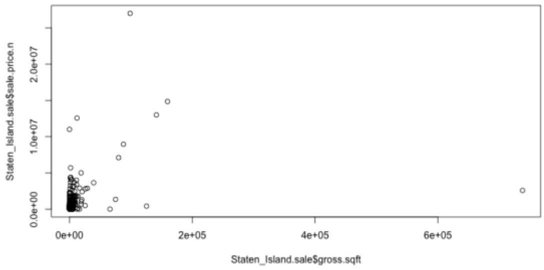
The Analysis of Staten_Island

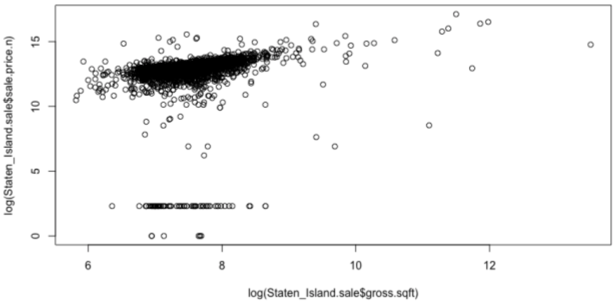




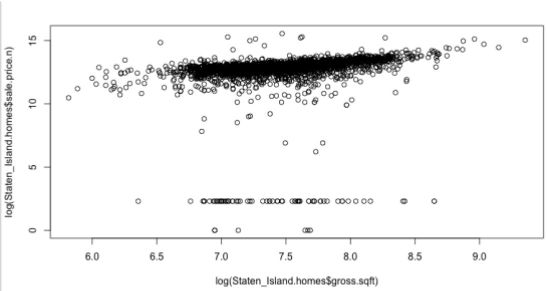


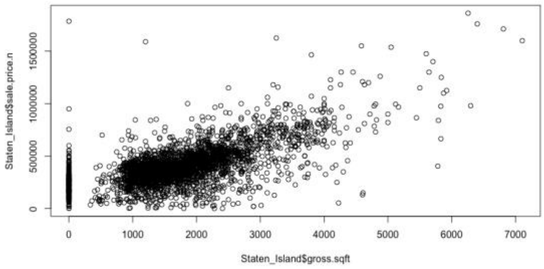
- *The Histogram of sale.price.n chart shows the data_Staten_Island has 5356 rows*
- *The Histogram of sale.price.n[sale.price.n>0] shows the sale price large than 0 are 2777 rows which means there are 2579 price value are less than 0.*



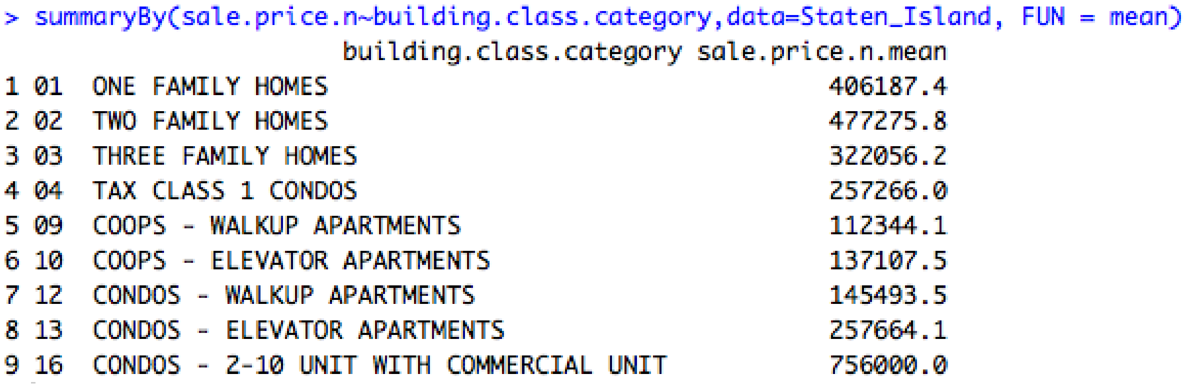


- *Only leave the actual sales only and show the plot of gross sqft and sale price. There still some points that the gross sqft is high but the price is low or the gross sqft is small but the sale price is high.*
- *There are several outliers which influence the distribution of the plot.*
- *Log them and see the relationship between these two variables.*

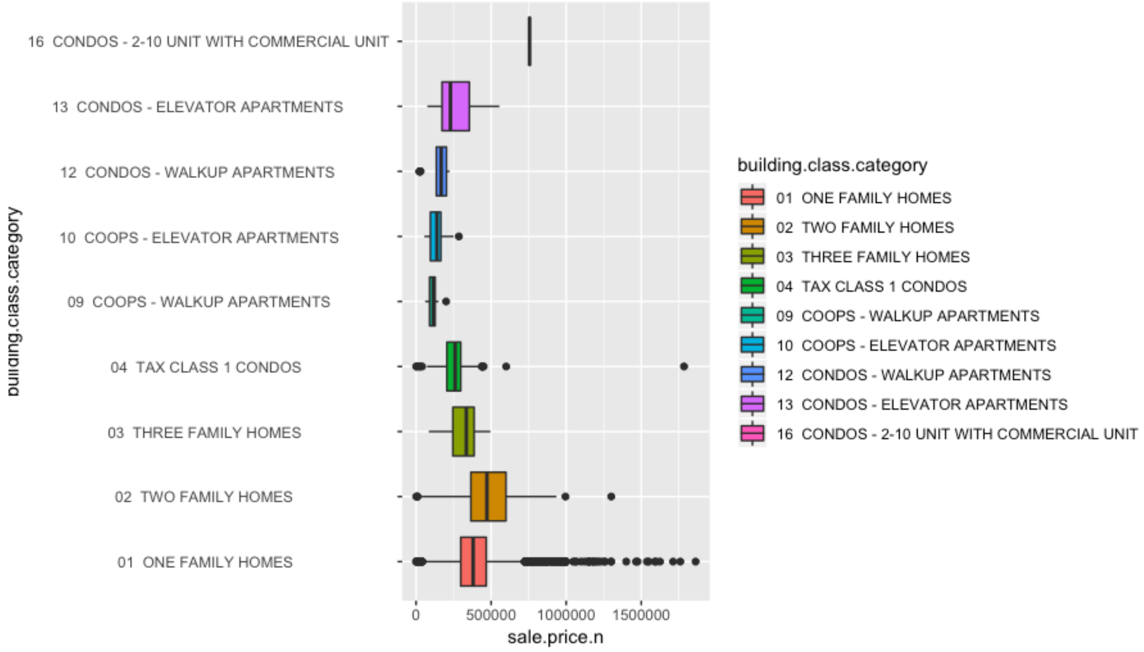




- *From above left chart we can make sure where is the outliers. We need to remove value which the log(Staten_Island.homes$gross.sqft) >10 and the log(Staten_Island.homes$sale.price.n) < 5, Which means the Staten_Island.homes$gross.sqft > 8000 and the Staten_Island.homes$sale.price.n < 100.*
- *The above right chart is the Staten_Island data after clean up. But there must be a lot of missing gross sqft in the data set. The gross.sqft is 0 but have the sale price.*



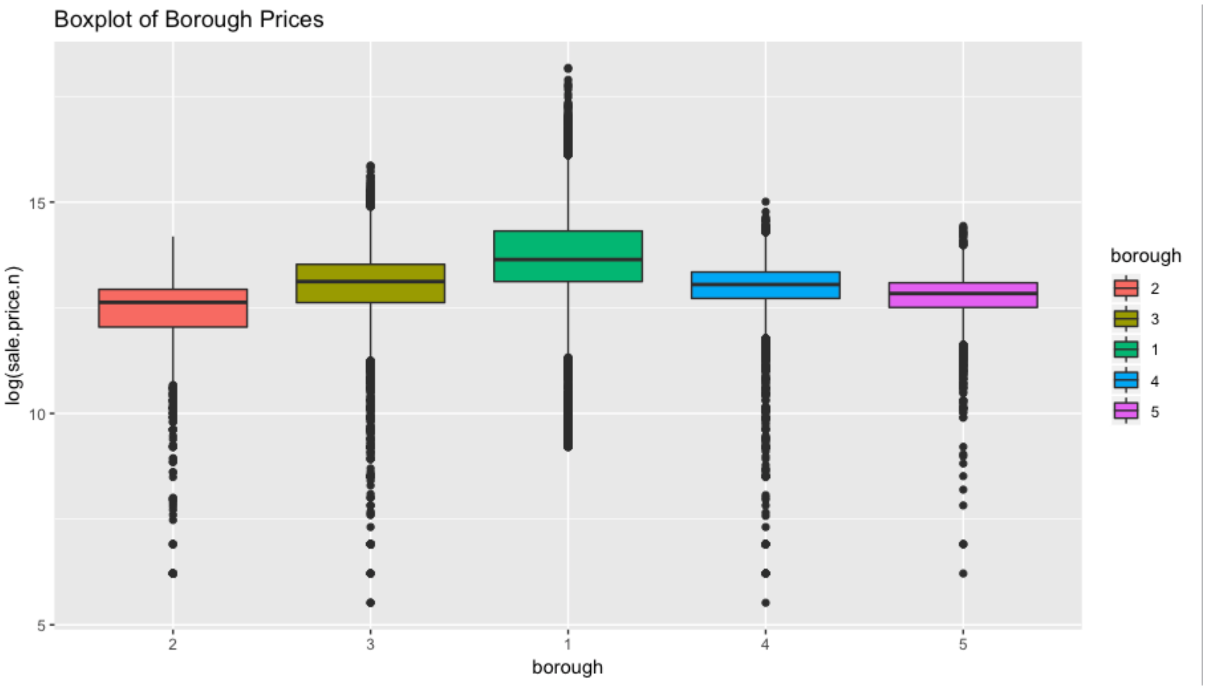
```
> summaryBy(sale.price.n~building.class.category,data=Staten_Island, FUN = mean)
                           building.class.category sale.price.n.mean
1 01  ONE FAMILY HOMES                                      406187.4
2 02  TWO FAMILY HOMES                                      477275.8
3 03  THREE FAMILY HOMES                                    322056.2
4 04  TAX CLASS 1 CONDOS                                    257266.0
5 09  COOPS - WALKUP APARTMENTS                             112344.1
6 10  COOPS - ELEVATOR APARTMENTS                           137107.5
7 12  CONDOS - WALKUP APARTMENTS                            145493.5
8 13  CONDOS - ELEVATOR APARTMENTS                          257664.1
9 16  CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT               756000.0
```

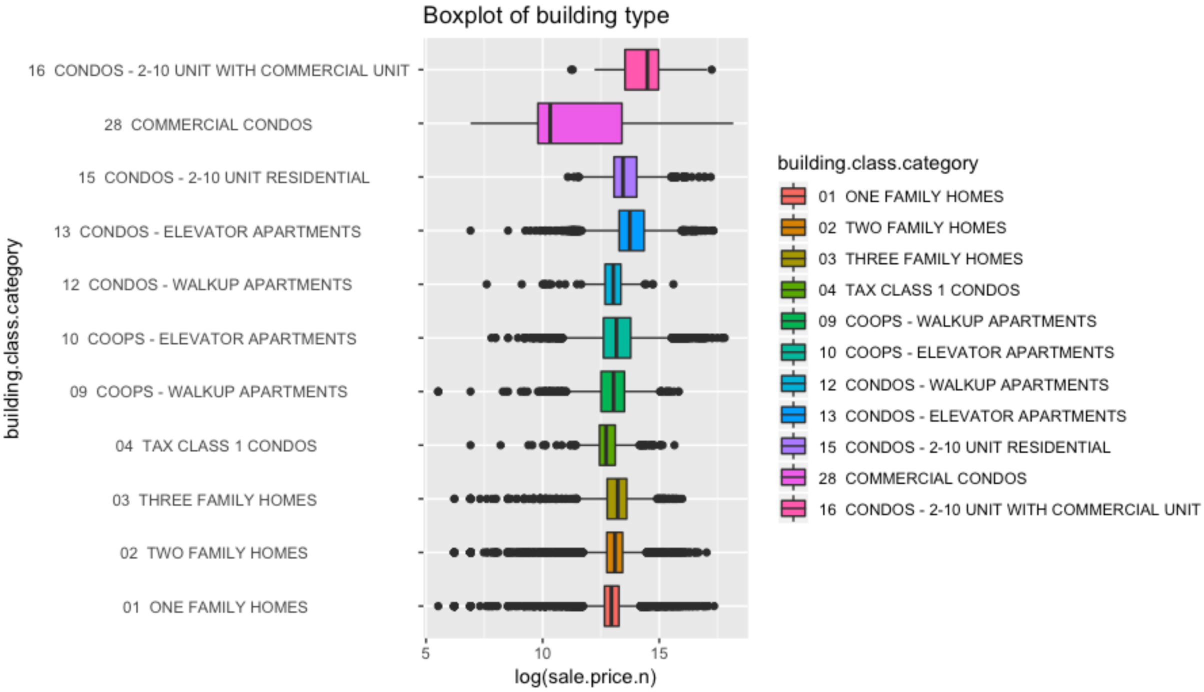


Here is the sale price of different building type boxplot in Staten_Island

- *In Staten_Island the two family mean price is the largest one except 16 Condos.*
- *The one family homes have the largest range.*
- *The 16 condos have too small samples so the mean price value is extremely high than the others building price.*

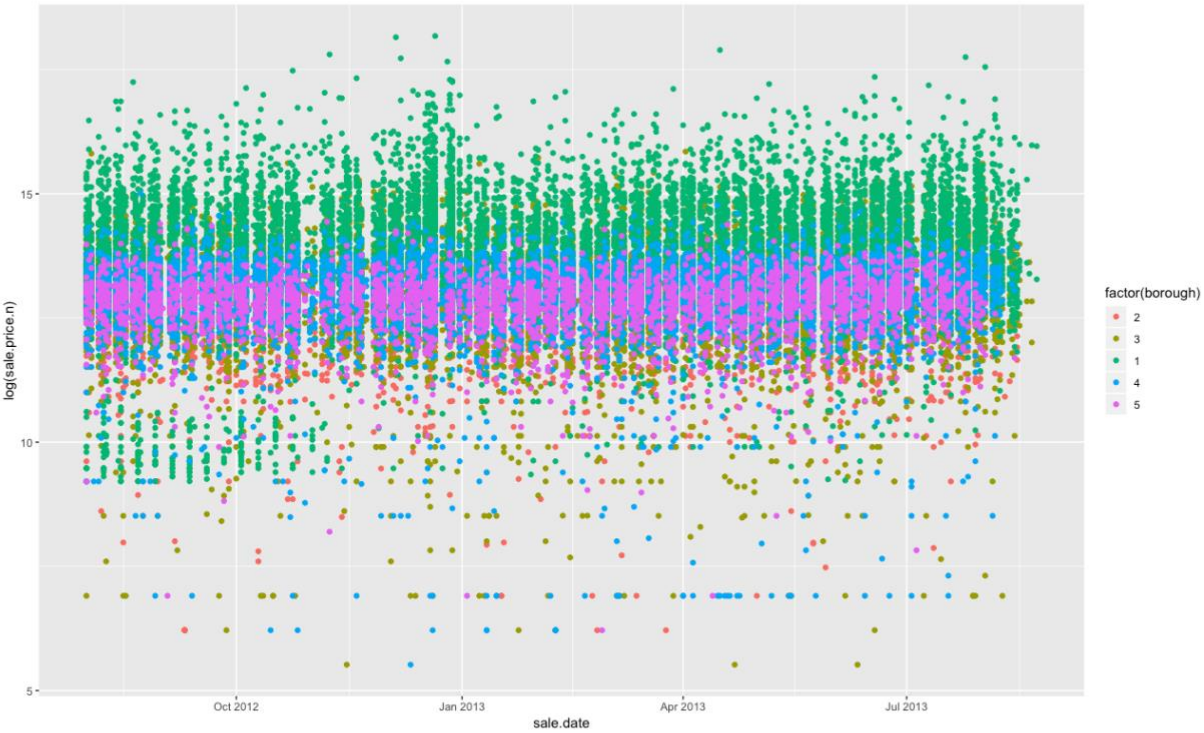
Combine five dataset





Conclusion:

- Overall the Commercial Condos have the low mean sale price but large range price. 0nly Manhattan, Brooklyn, Bronx have this building type. And the price of Brooklyn and Bronx is the lowest one but this type of building have the largest price range and have a lot of extremely high price values.
- The Borough 1 which is Manhattan is higher than other boroughs. So the Manhattan is the most expensive area.



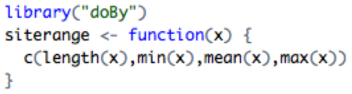
Time Price chart

Conclusion:

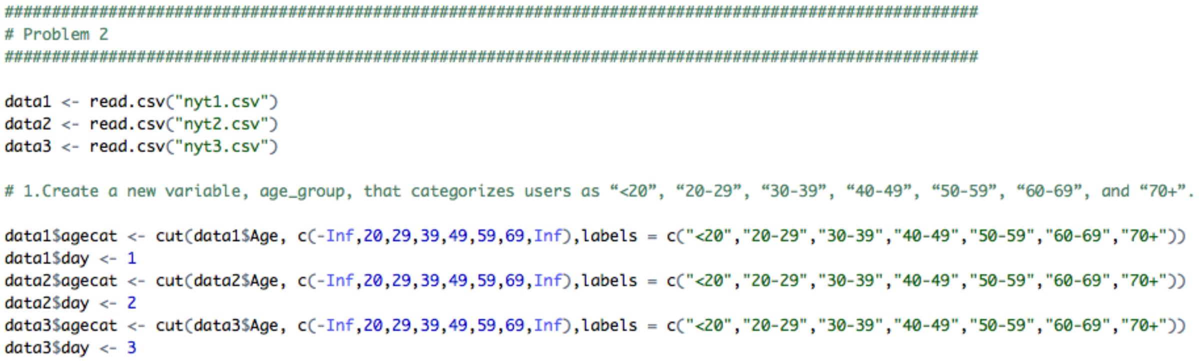
- *As the plot show above there is no specific trend tHrough the sale date with the price. Overall borough 1 which is Manhattan is the most expensive one.*

Problem 2

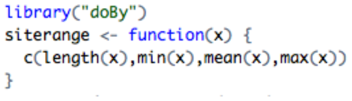
- *Create a new variable, age_group, that categorizes users as "<20", "20-29", "30-39", "40-49", "50-59", "60-69", and "70+".*



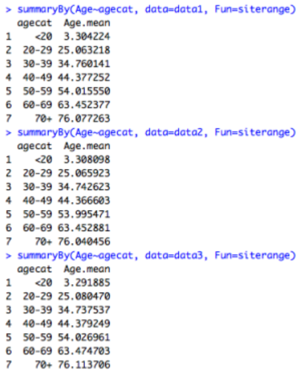
```
library("doBy")
siterange <- function(x) {
  c(length(x),min(x),mean(x),max(x))
}
```



```
##########################################################################################################
# Problem 2
##########################################################################################################

data1 <- read.csv("nyt1.csv")
data2 <- read.csv("nyt2.csv")
data3 <- read.csv("nyt3.csv")

# 1.Create a new variable, age_group, that categorizes users as "<20", "20-29", "30-39", "40-49", "50-59", "60-69", and "70+".

data1$agecat <- cut(data1$Age, c(-Inf,20,29,39,49,59,69,Inf),labels = c("<20","20-29","30-39","40-49","50-59","60-69","70+"))
data1$day <- 1
data2$agecat <- cut(data2$Age, c(-Inf,20,29,39,49,59,69,Inf),labels = c("<20","20-29","30-39","40-49","50-59","60-69","70+"))
data2$day <- 2
data3$agecat <- cut(data3$Age, c(-Inf,20,29,39,49,59,69,Inf),labels = c("<20","20-29","30-39","40-49","50-59","60-69","70+"))
data3$day <- 3
```

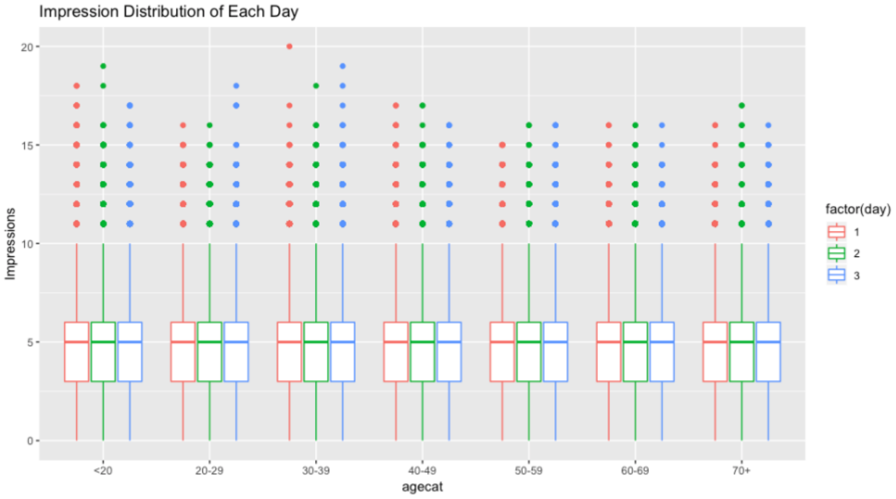


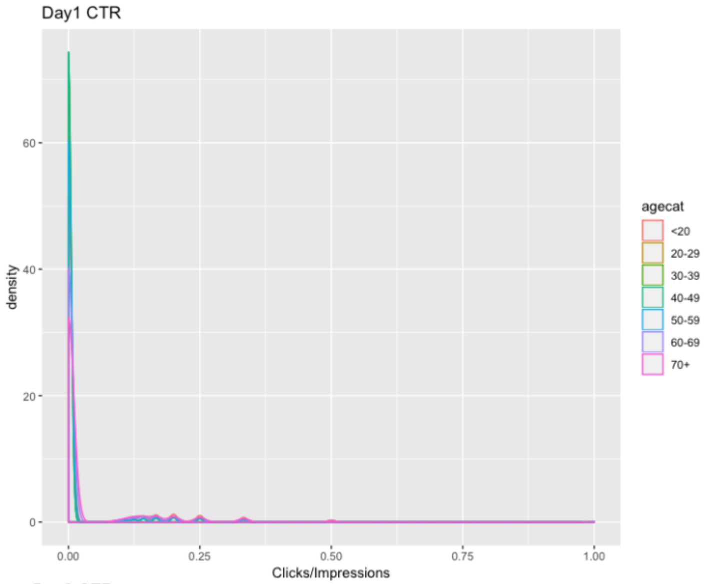
```
library("doBy")
siterange <- function(x) {
  c(length(x),min(x),mean(x),max(x))
}
```

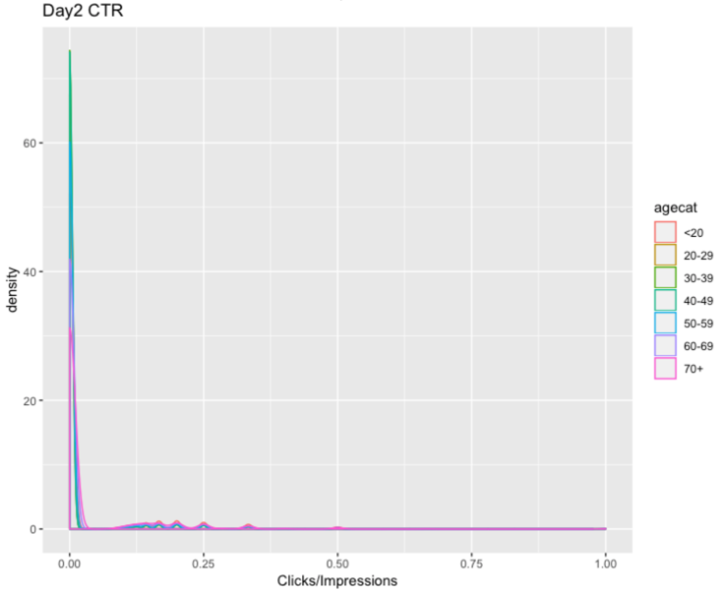


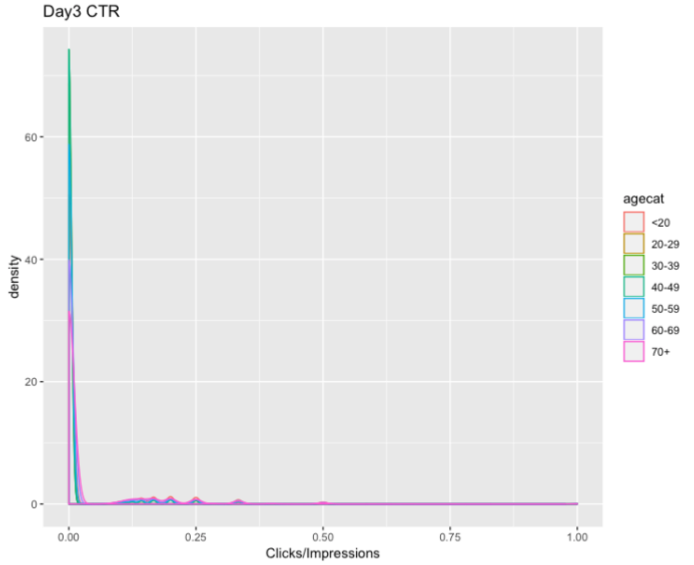
```
> summaryBy(Age~agecat, data=data1, Fun=siterange)
  agecat  Age.mean
1    <20  3.304224
2  20-29 25.063218
3  30-39 34.760141
4  40-49 44.377252
5  50-59 54.015550
6  60-69 63.452377
7    70+ 76.077263
> summaryBy(Age~agecat, data=data2, Fun=siterange)
  agecat  Age.mean
1    <20  3.308098
2  20-29 25.065923
3  30-39 34.742623
4  40-49 44.366603
5  50-59 53.995471
6  60-69 63.452881
7    70+ 76.040456
> summaryBy(Age~agecat, data=data3, Fun=siterange)
  agecat  Age.mean
1    <20  3.291885
2  20-29 25.080470
3  30-39 34.737537
4  40-49 44.379249
5  50-59 54.026961
6  60-69 63.474703
7    70+ 76.113706
```

- *For each day:*
  - *o Plot the distribution of number of impressions and click-through-rate (CTR = #clicks / #impressions) for these age categories*

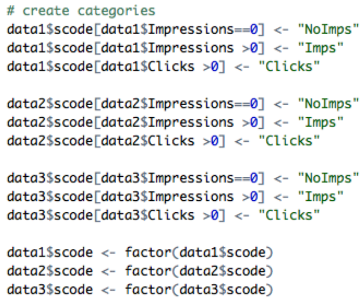



Day1 CTR
density
0
20
40
60
0.00
0.25
0.50
0.75
1.00
Clicks/Impressions
agecat
<20
20-29
30-39
40-49
50-59
60-69
70+


Day2 CTR
density
0
20
40
60
0.00
0.25
0.50
0.75
1.00
Clicks/Impressions
agecat
<20
20-29
30-39
40-49
50-59
60-69
70+


Day3 CTR
density
0
20
40
60
0.00
0.25
0.50
0.75
1.00
Clicks/Impressions
agecat
<20
20-29
30-39
40-49
50-59
60-69
70+

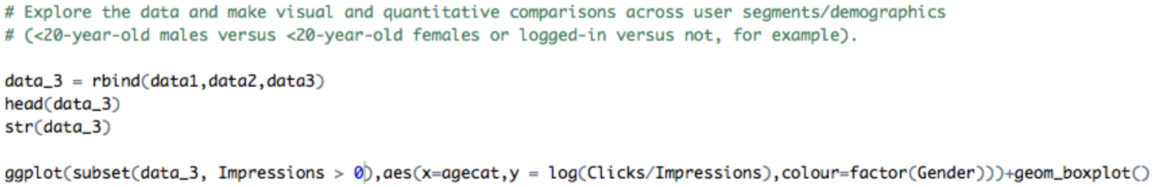
o *Define a new variable to segment or categorize users based on their click behavior.*



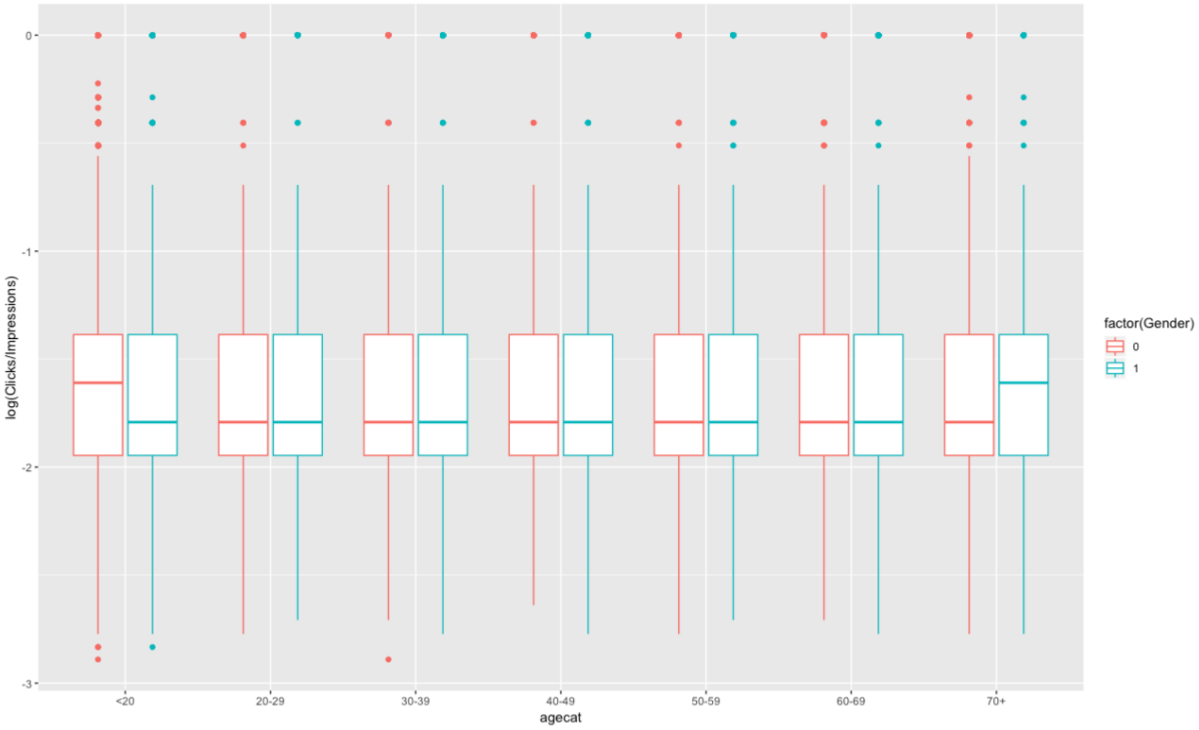
```
# create categories
data1$scode[data1$Impressions==0] <- "NoImps"
data1$scode[data1$Impressions >0] <- "Imps"
data1$scode[data1$Clicks >0] <- "Clicks"

data2$scode[data2$Impressions==0] <- "NoImps"
data2$scode[data2$Impressions >0] <- "Imps"
data2$scode[data2$Clicks >0] <- "Clicks"

data3$scode[data3$Impressions==0] <- "NoImps"
data3$scode[data3$Impressions >0] <- "Imps"
data3$scode[data3$Clicks >0] <- "Clicks"

data1$scode <- factor(data1$scode)
data2$scode <- factor(data2$scode)
data3$scode <- factor(data3$scode)
```

o *Explore the data and make visual and quantitative comparisons across user segments/demographics (<20-year-old males versus <20-year-old females or logged-in versus not, for example).*

What the difference of CTR between different age and different gender?



```
# Explore the data and make visual and quantitative comparisons across user segments/demographics
# (<20-year-old males versus <20-year-old females or logged-in versus not, for example).

data_3 = rbind(data1,data2,data3)
head(data_3)
str(data_3)

ggplot(subset(data_3, Impressions > 0),aes(x=agecat,y = log(Clicks/Impressions),colour=factor(Gender)))+geom_boxplot()
```



0 = female. 1= male

Conclusion:

As above chart shows, the mean CTR of under20 female are higher than the under20 male mean CTR of these total tree days. The male older than 70 have higher CTR than older than 70 females’. The other age do not have significant different from the box plot.