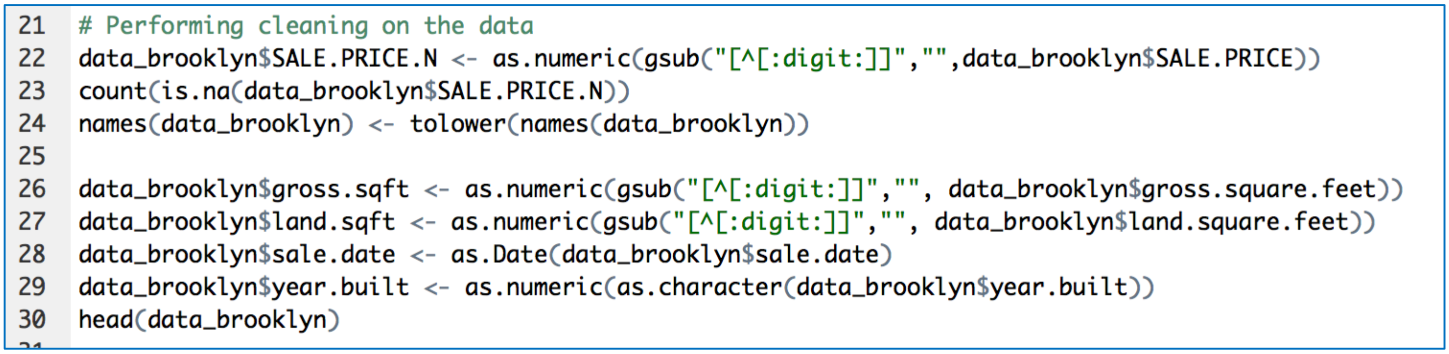
DATA INTENSIVE COMPUTING

Data Economy: A Real Case Study

Ramanpreet Singh Khinda | DIC 587 | March 5, 2016
rkhinda@buffalo.edu
50169622

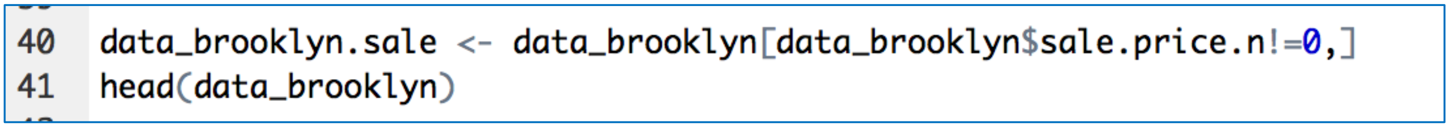
Problem 3A: EDA on Brooklyn Rolling Sales data

Step 1: Cleaning the data and Performing EDA

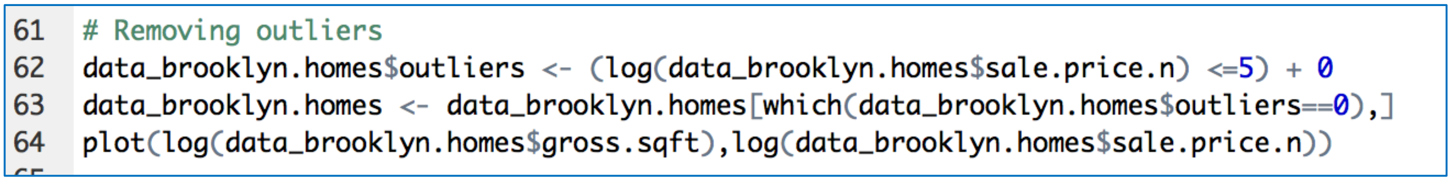


```r
# Performing cleaning on the data
data_brooklyn$SALE.PRICE.N <- as.numeric(gsub("[^[:digit:]]","",data_brooklyn$SALE.PRICE))
count(is.na(data_brooklyn$SALE.PRICE.N))
names(data_brooklyn) <- tolower(names(data_brooklyn))

data_brooklyn$gross.sqft <- as.numeric(gsub("[^[:digit:]]","", data_brooklyn$gross.square.feet))
data_brooklyn$land.sqft <- as.numeric(gsub("[^[:digit:]]","", data_brooklyn$land.square.feet))
data_brooklyn$sale.date <- as.Date(data_brooklyn$sale.date)
data_brooklyn$year.built <- as.numeric(as.character(data_brooklyn$year.built))
head(data_brooklyn)
```

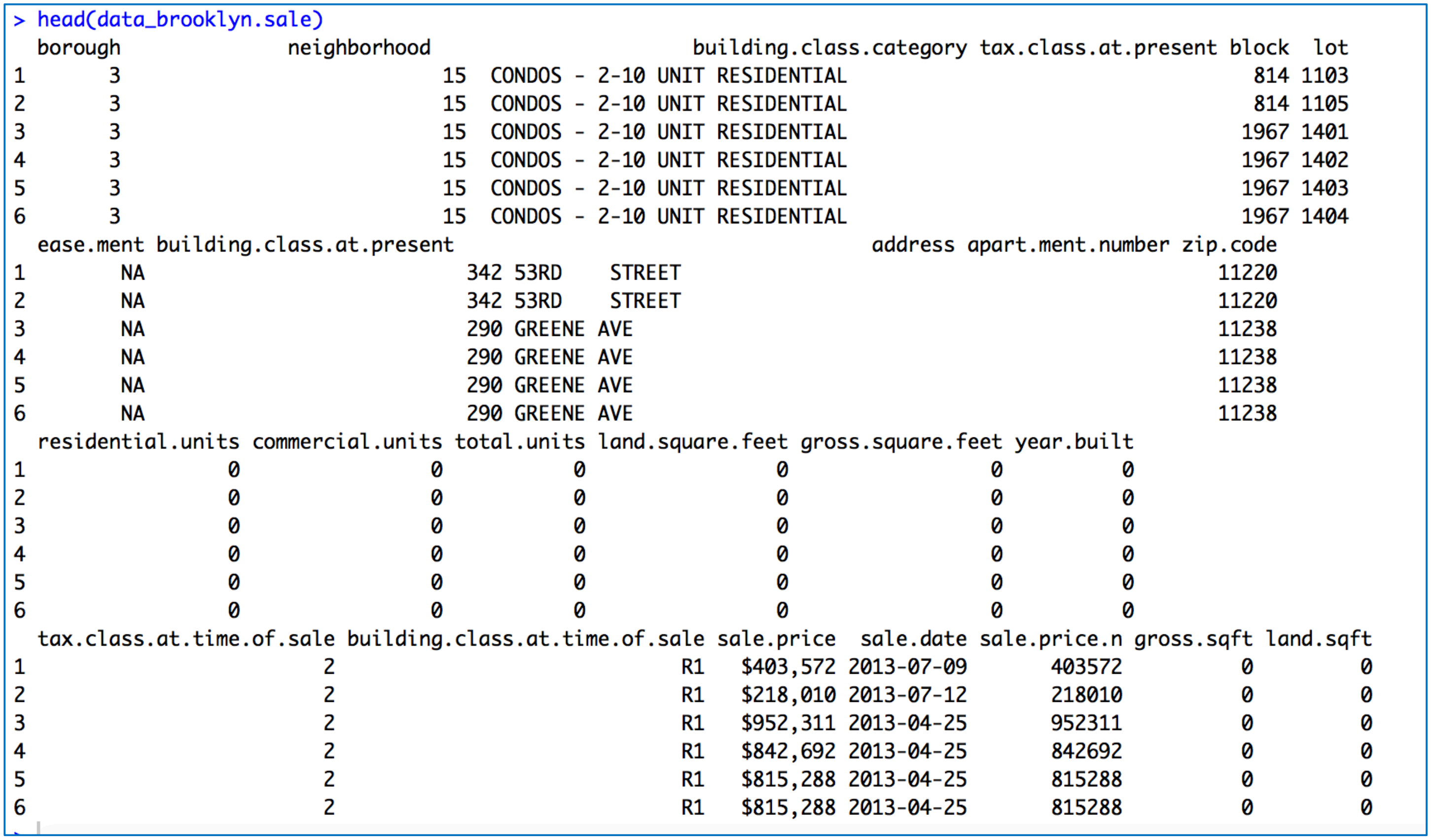


```r
data_brooklyn.sale <- data_brooklyn[data_brooklyn$sale.price.n!=0,]
head(data_brooklyn)
```

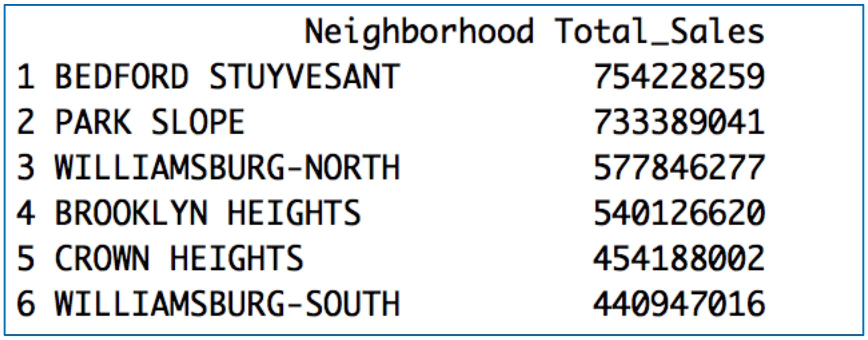


```r
# Removing outliers
data_brooklyn.homes$outliers <- (log(data_brooklyn.homes$sale.price.n) <=5) + 0
data_brooklyn.homes <- data_brooklyn.homes[which(data_brooklyn.homes$outliers==0),]
plot(log(data_brooklyn.homes$gross.sqft),log(data_brooklyn.homes$sale.price.n))
```

After cleaning The data looks something like this

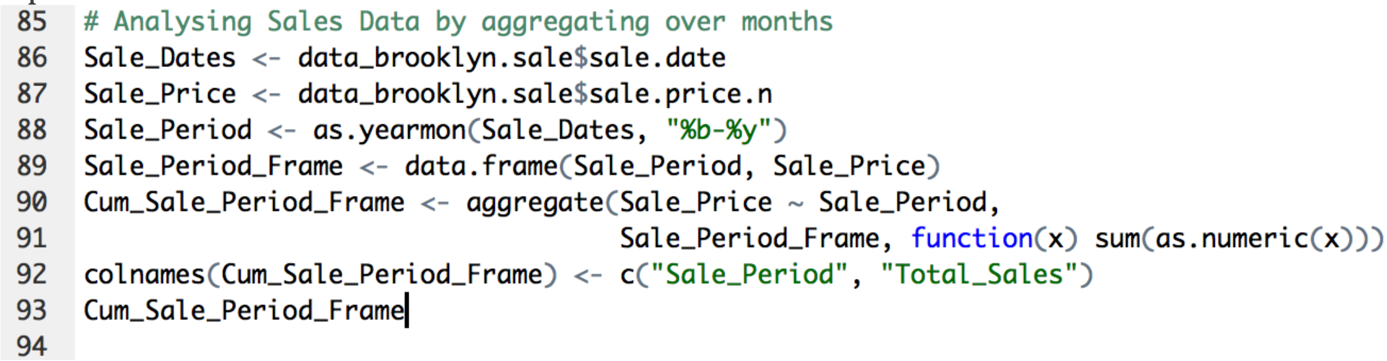


```
> head(data_brooklyn.sale)
  borough              neighborhood                      building.class.category tax.class.at.present block  lot
1       3                           15  CONDOS - 2-10 UNIT RESIDENTIAL                                  814 1103
2       3                           15  CONDOS - 2-10 UNIT RESIDENTIAL                                  814 1105
3       3                           15  CONDOS - 2-10 UNIT RESIDENTIAL                                 1967 1401
4       3                           15  CONDOS - 2-10 UNIT RESIDENTIAL                                 1967 1402
5       3                           15  CONDOS - 2-10 UNIT RESIDENTIAL                                 1967 1403
6       3                           15  CONDOS - 2-10 UNIT RESIDENTIAL                                 1967 1404
  ease.ment building.class.at.present                                   address apart.ment.number zip.code
1        NA                           342 53RD    STREET                                             11220
2        NA                           342 53RD    STREET                                             11220
3        NA                           290 GREENE AVE                                                 11238
4        NA                           290 GREENE AVE                                                 11238
5        NA                           290 GREENE AVE                                                 11238
6        NA                           290 GREENE AVE                                                 11238
  residential.units commercial.units total.units land.square.feet gross.square.feet year.built
1                 0                0           0                0                 0          0
2                 0                0           0                0                 0          0
3                 0                0           0                0                 0          0
4                 0                0           0                0                 0          0
5                 0                0           0                0                 0          0
6                 0                0           0                0                 0          0
  tax.class.at.time.of.sale building.class.at.time.of.sale sale.price  sale.date sale.price.n gross.sqft land.sqft
1                         2                             R1   $403,572 2013-07-09       403572          0         0
2                         2                             R1   $218,010 2013-07-12       218010          0         0
3                         2                             R1   $952,311 2013-04-25       952311          0         0
4                         2                             R1   $842,692 2013-04-25       842692          0         0
5                         2                             R1   $815,288 2013-04-25       815288          0         0
6                         2                             R1   $815,288 2013-04-25       815288          0         0
```



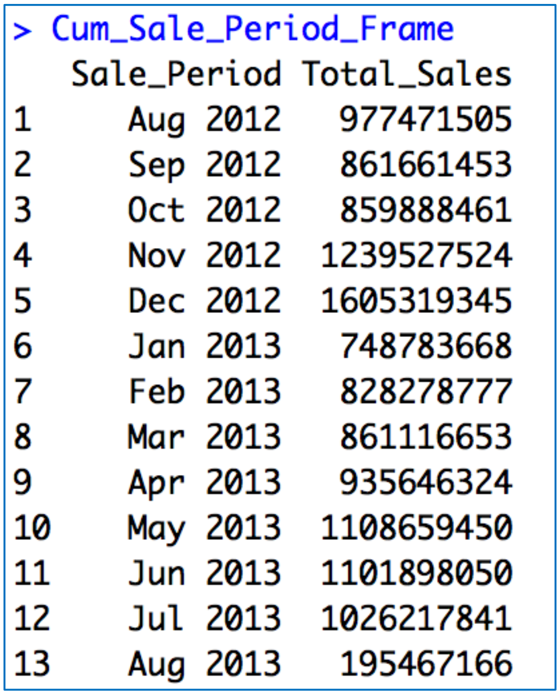
```
            Neighborhood Total_Sales
1 BEDFORD STUYVESANT       754228259
2 PARK SLOPE               733389041
3 WILLIAMSBURG-NORTH       577846277
4 BROOKLYN HEIGHTS         540126620
5 CROWN HEIGHTS            454188002
6 WILLIAMSBURG-SOUTH       440947016
```

We also performed aggregation on Sales Period so as to get intuition on the Total Sales for a particular month



```r
# Analysing Sales Data by aggregating over months
Sale_Dates <- data_brooklyn.sale$sale.date
Sale_Price <- data_brooklyn.sale$sale.price.n
Sale_Period <- as.yearmon(Sale_Dates, "%b-%y")
Sale_Period_Frame <- data.frame(Sale_Period, Sale_Price)
Cum_Sale_Period_Frame <- aggregate(Sale_Price ~ Sale_Period,
                                  Sale_Period_Frame, function(x) sum(as.numeric(x)))
colnames(Cum_Sale_Period_Frame) <- c("Sale_Period", "Total_Sales")
Cum_Sale_Period_Frame
```

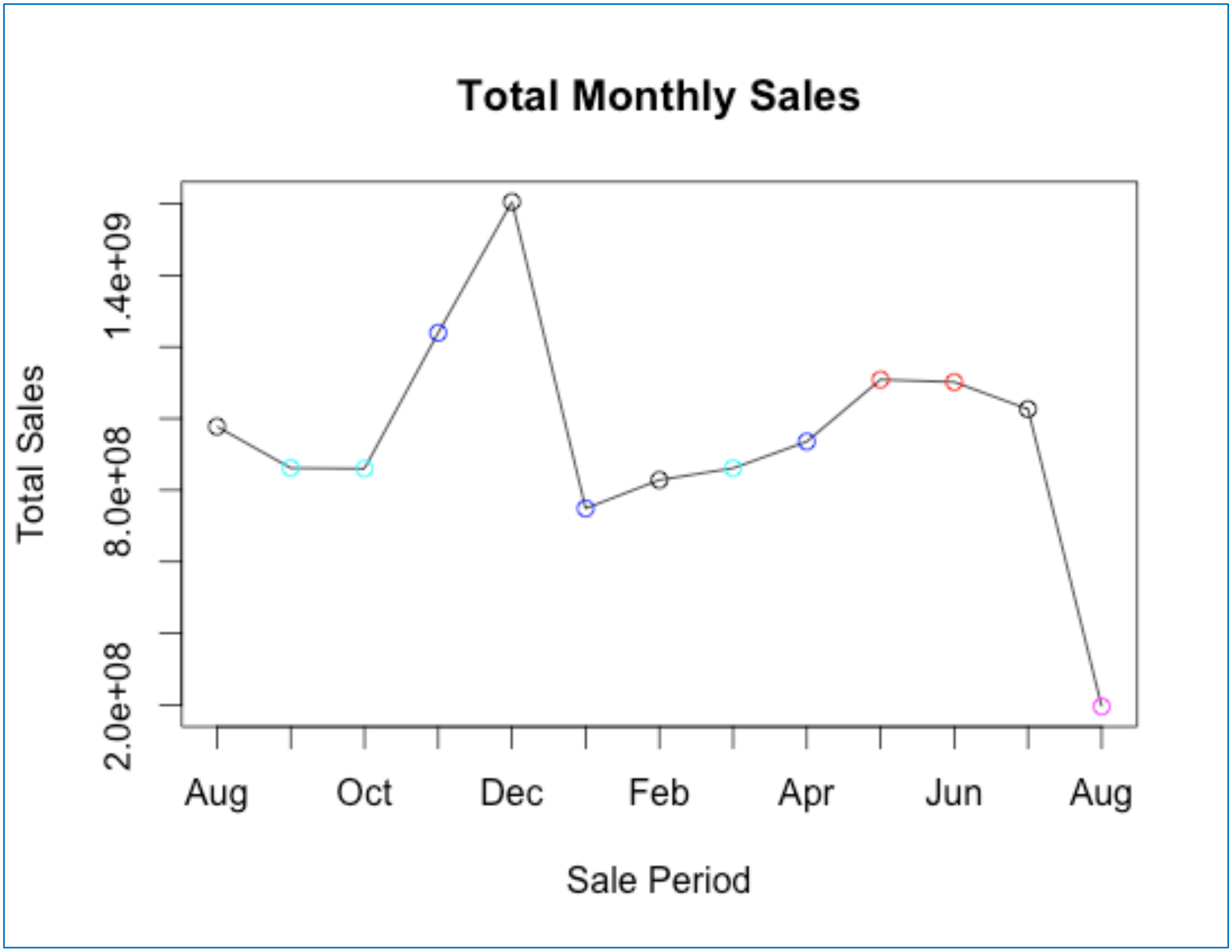
Cumulative Total Sales over the year

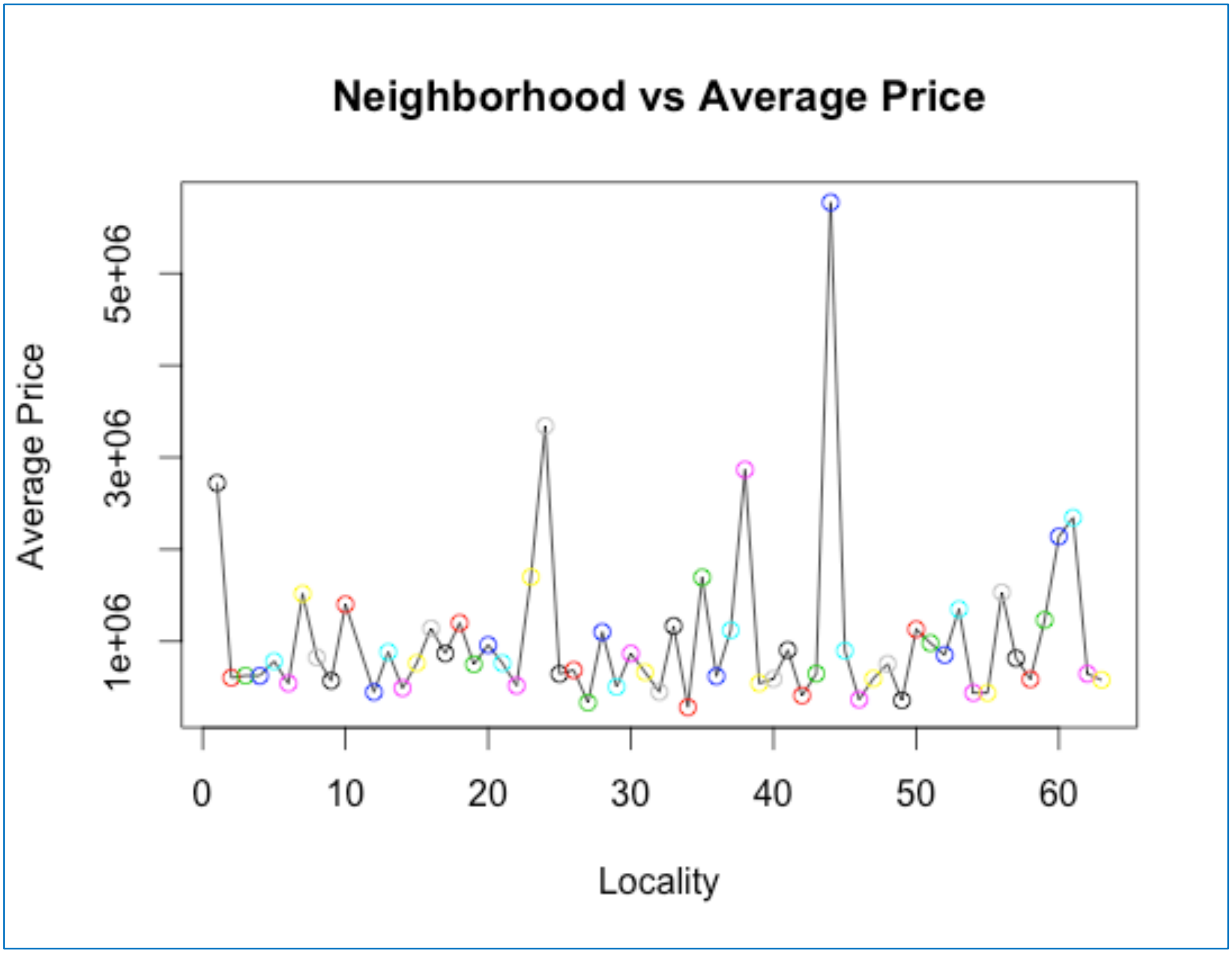


```
> Cum_Sale_Period_Frame
   Sale_Period Total_Sales
1     Aug 2012   977471505
2     Sep 2012   861661453
3     Oct 2012   859888461
4     Nov 2012  1239527524
5     Dec 2012  1605319345
6     Jan 2013   748783668
7     Feb 2013   828278777
8     Mar 2013   861116653
9     Apr 2013   935646324
10    May 2013  1108659450
11    Jun 2013  1101898050
12    Jul 2013  1026217841
13    Aug 2013   195467166
```

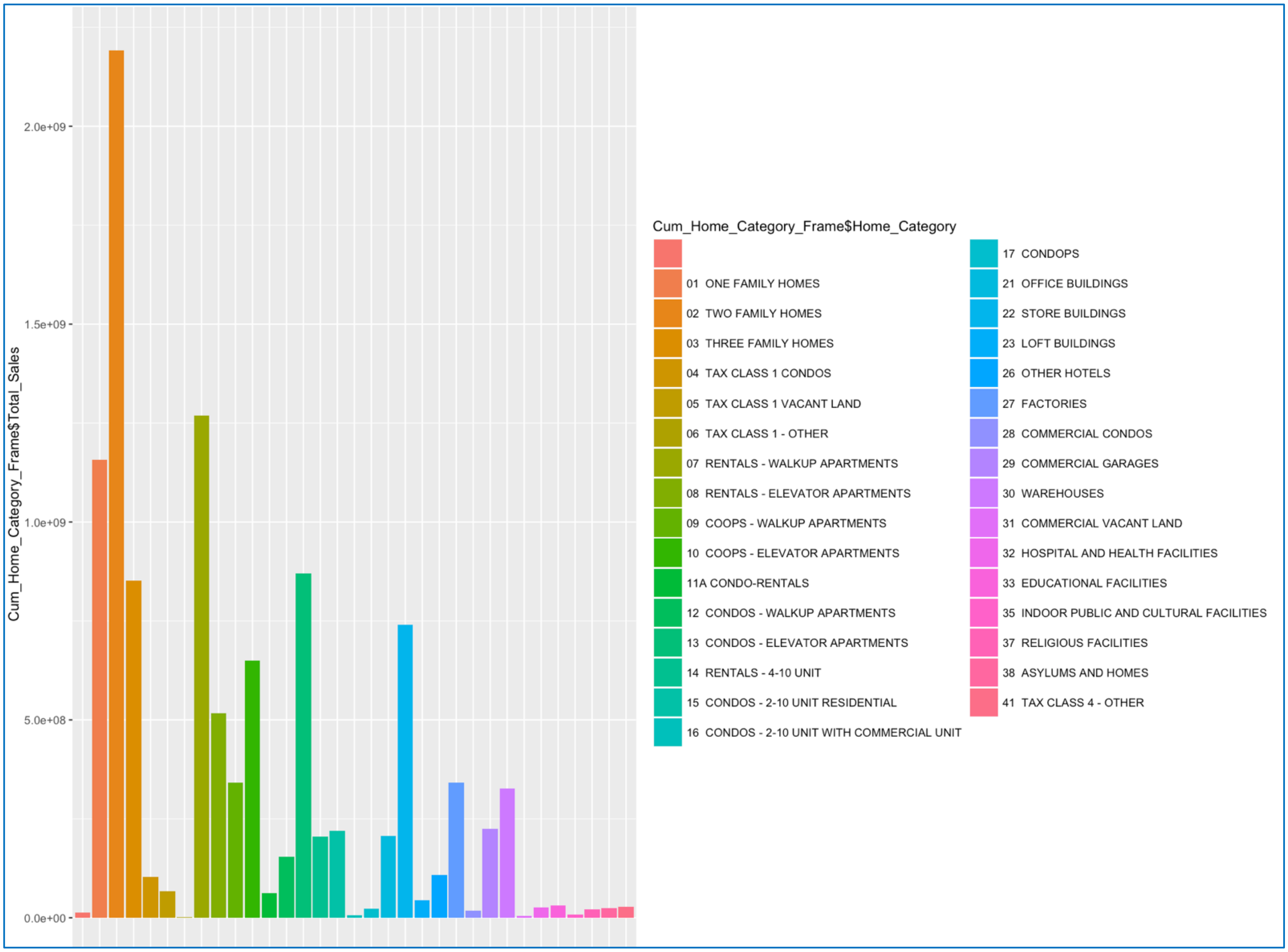
Step 2: Generating graphs

The next step is to gather meaning information on the sales for the Brooklyn by generating different graphs and plots




Neighborhood vs Average Price
Average Price
1e+06
3e+06
5e+06
0
10
20
30
40
50
60
Locality

Graph for Cumulative Total Sales over the year vs Home Category



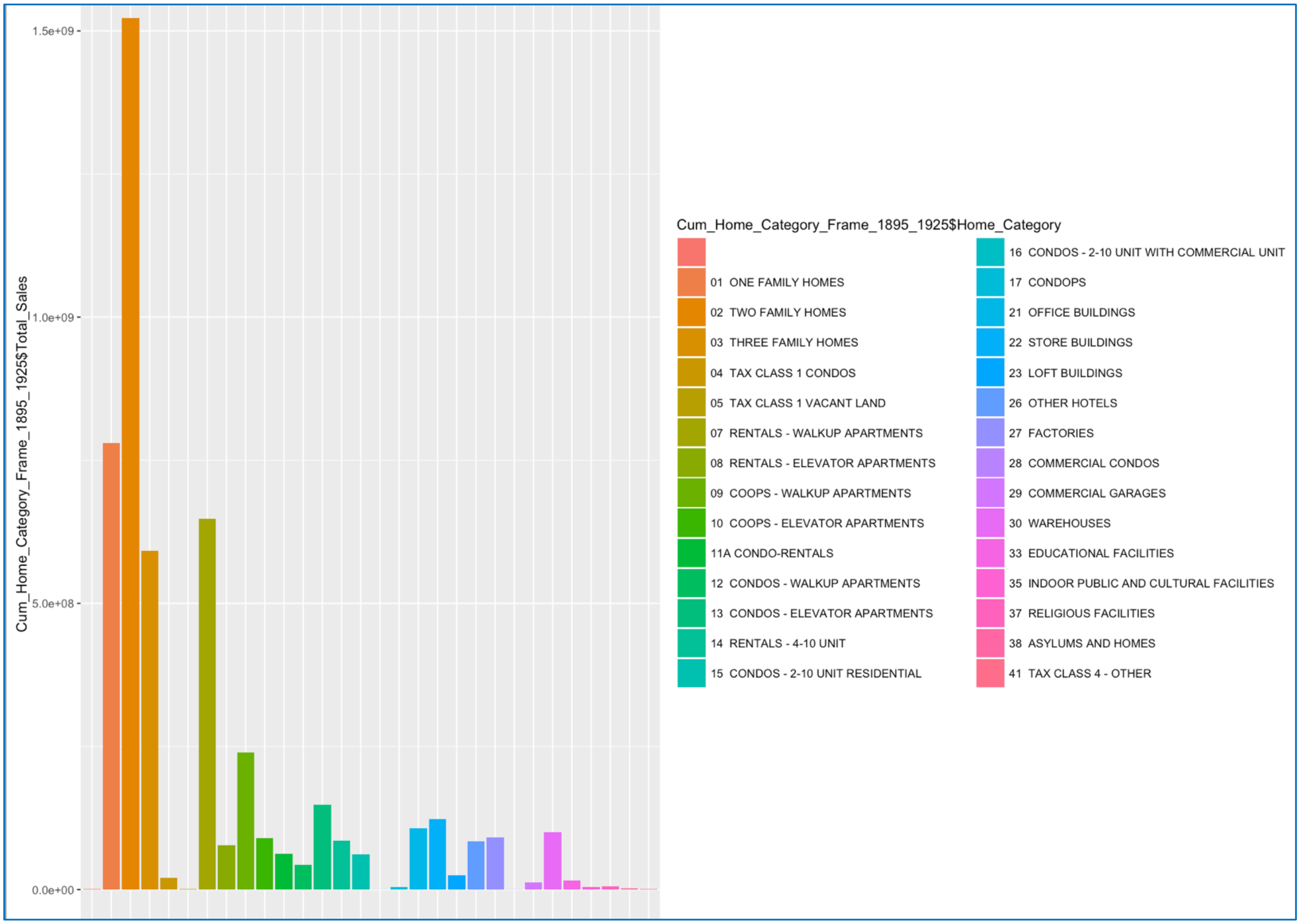
Step 3: Analysis in details

Now we take a step ahead and split the Home built dates into groups of 30 years each and analyze how the sales of Brooklyn are affecting over the years.

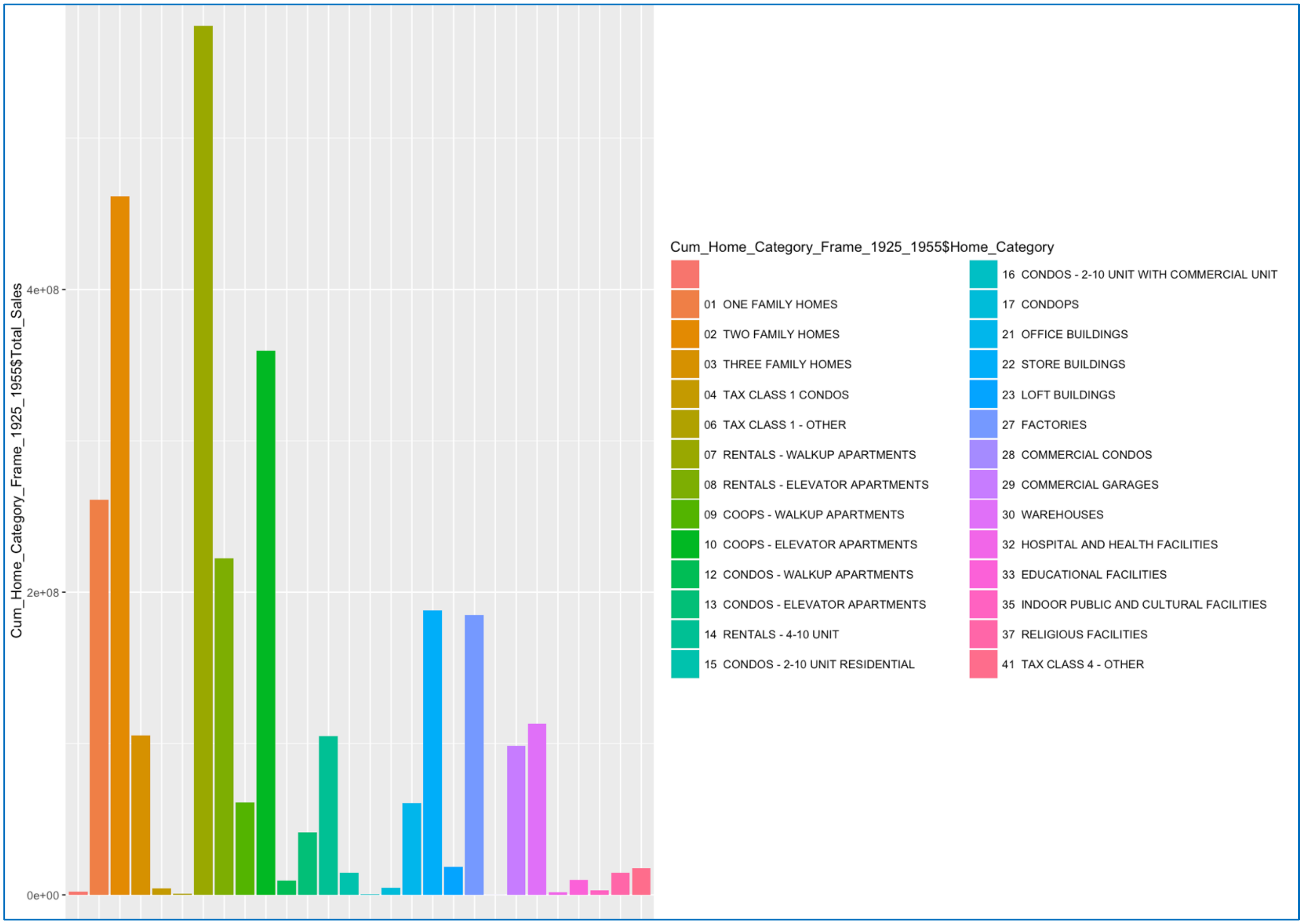
This will also give us intuition that which home categories got popular during these different time frames

Graphs for Cumulative Total Sales over the year vs Home Category

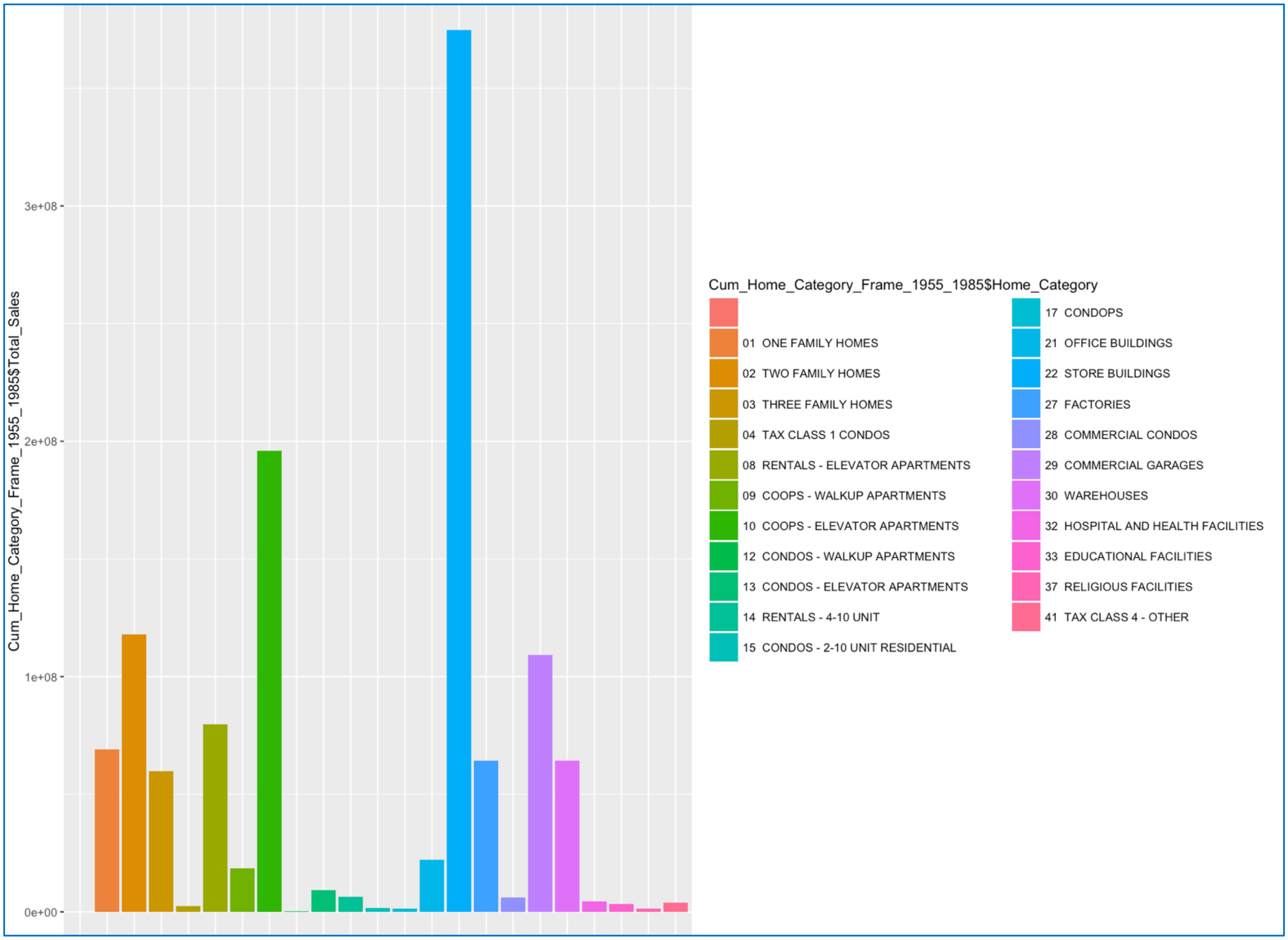
Period 1895 - 1925



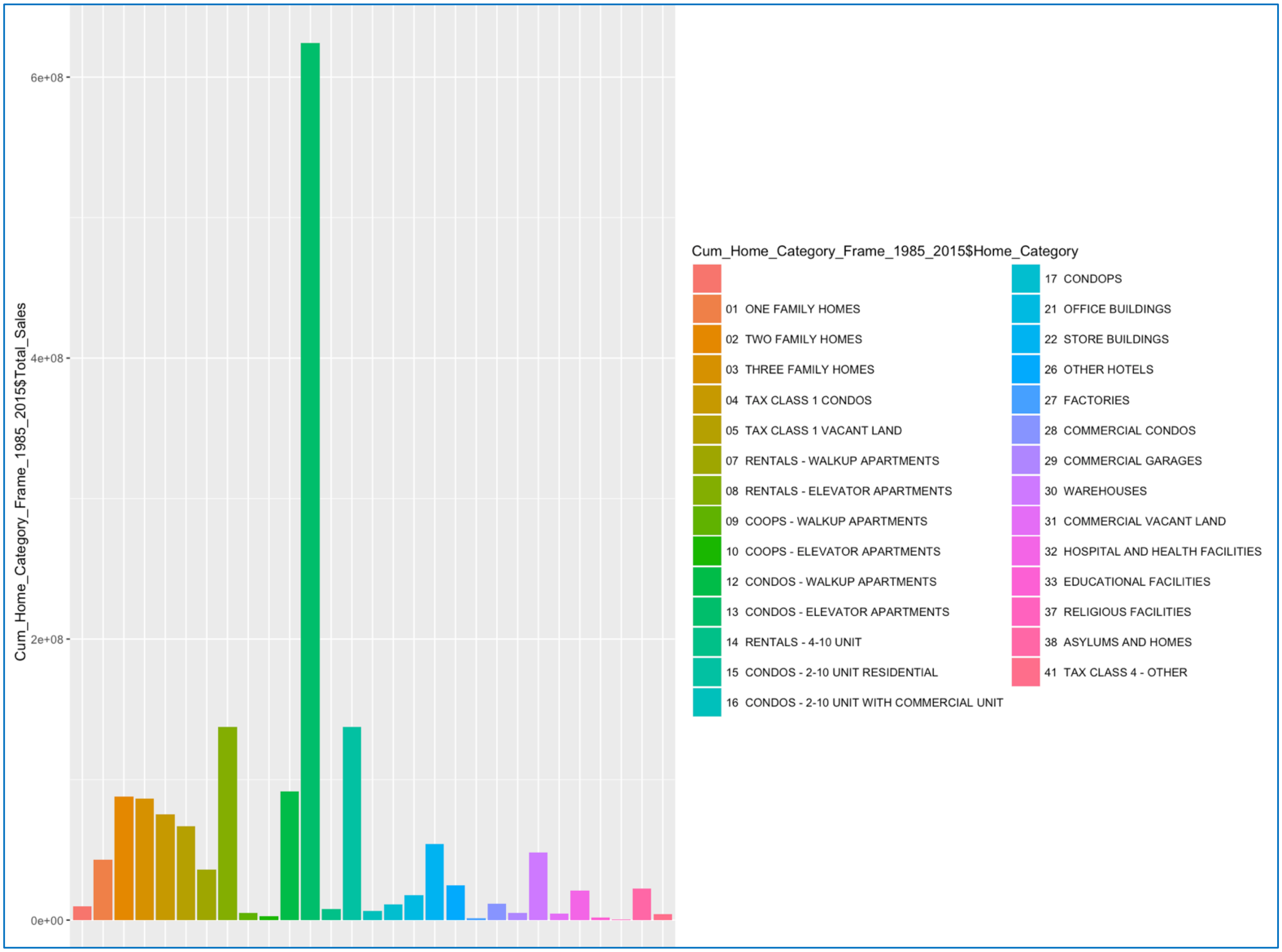
Period 1925-1955



Period 1955-1985



Period 1985-2015

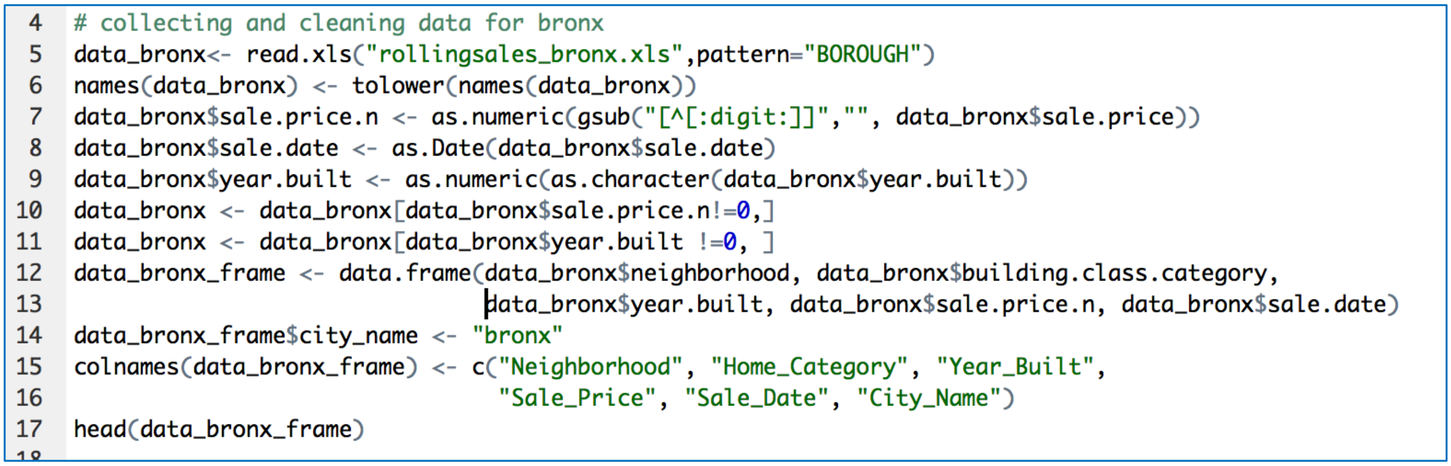


Problem 3B: EDA on data for Manhattan, Queens, Bronx and State Island

In this problem we extend our analysis on data for other locations. We

Step 1:

In the first step we will do cleaning on data. For example, the data cleaning process for Manhattan is described below. We will follow the same process for other locations as well

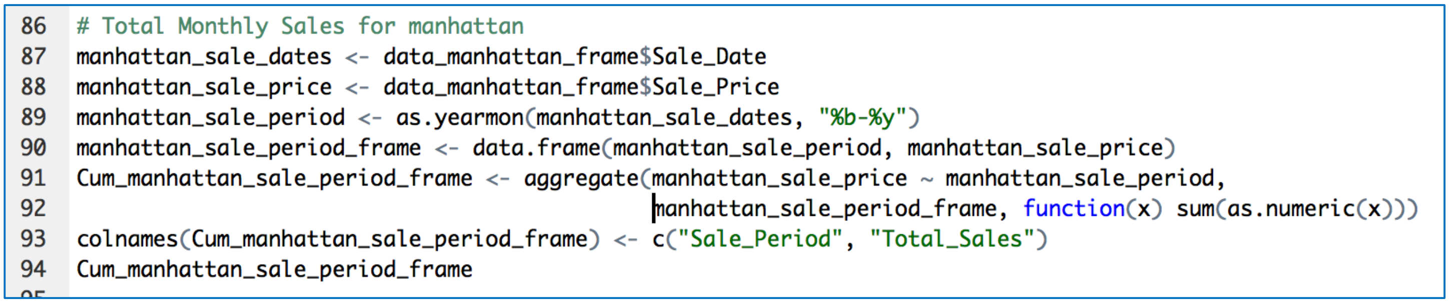


```
# collecting and cleaning data for bronx
data_bronx<- read.xls("rollingsales_bronx.xls",pattern="BOROUGH")
names(data_bronx) <- tolower(names(data_bronx))
data_bronx$sale.price.n <- as.numeric(gsub("[^[:digit:]]","", data_bronx$sale.price))
data_bronx$sale.date <- as.Date(data_bronx$sale.date)
data_bronx$year.built <- as.numeric(as.character(data_bronx$year.built))
data_bronx <- data_bronx[data_bronx$sale.price.n!=0,]
data_bronx <- data_bronx[data_bronx$year.built !=0, ]
data_bronx_frame <- data.frame(data_bronx$neighborhood, data_bronx$building.class.category,
                              data_bronx$year.built, data_bronx$sale.price.n, data_bronx$sale.date)
data_bronx_frame$city_name <- "bronx"
colnames(data_bronx_frame) <- c("Neighborhood", "Home_Category", "Year_Built",
                                "Sale_Price", "Sale_Date", "City_Name")
head(data_bronx_frame)
```

Step 2:

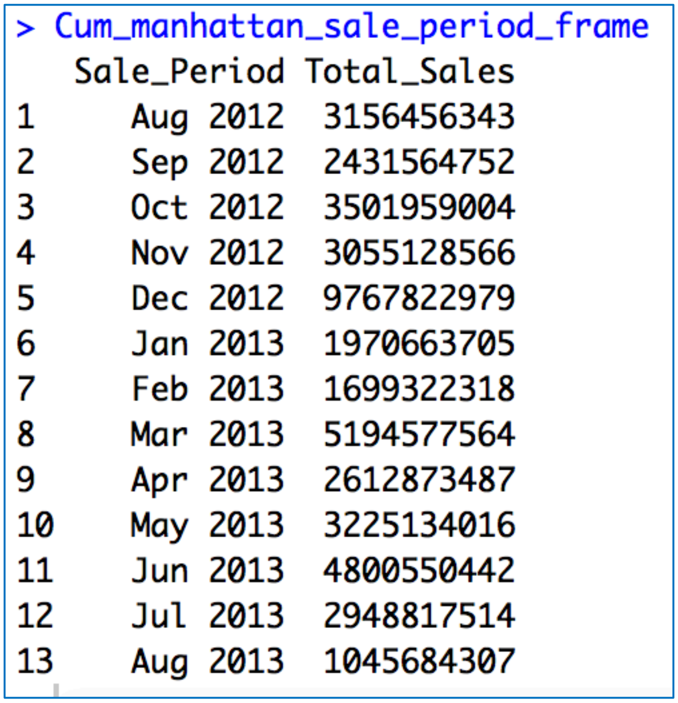
Now to analyze data for all the locations and compare them, we need to perform some aggregation so that its easy for us to do some analysis.

Since its sales data, so we decided to prepare cumulative sales report aggregated over the months for each location



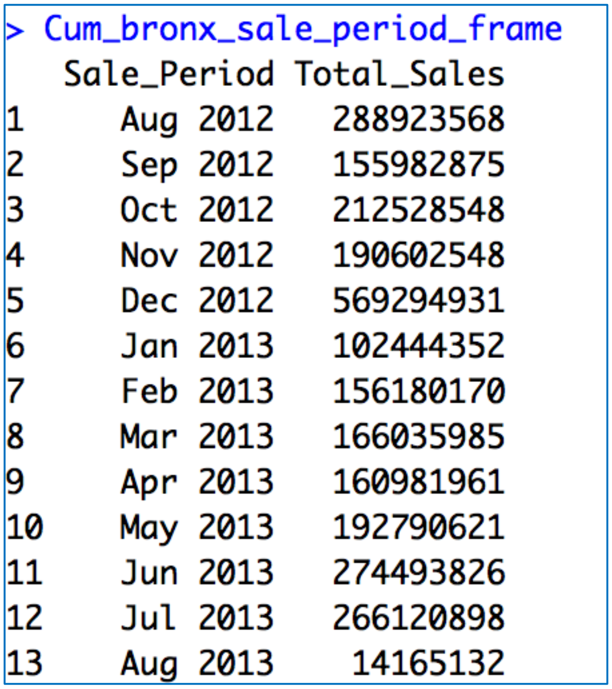
```
# Total Monthly Sales for manhattan
manhattan_sale_dates <- data_manhattan_frame$Sale_Date
manhattan_sale_price <- data_manhattan_frame$Sale_Price
manhattan_sale_period <- as.yearmon(manhattan_sale_dates, "%b-%y")
manhattan_sale_period_frame <- data.frame(manhattan_sale_period, manhattan_sale_price)
Cum_manhattan_sale_period_frame <- aggregate(manhattan_sale_price ~ manhattan_sale_period,
                                             manhattan_sale_period_frame, function(x) sum(as.numeric(x)))
colnames(Cum_manhattan_sale_period_frame) <- c("Sale_Period", "Total_Sales")
Cum_manhattan_sale_period_frame
```

Cumulative Sales Report of Manhattan



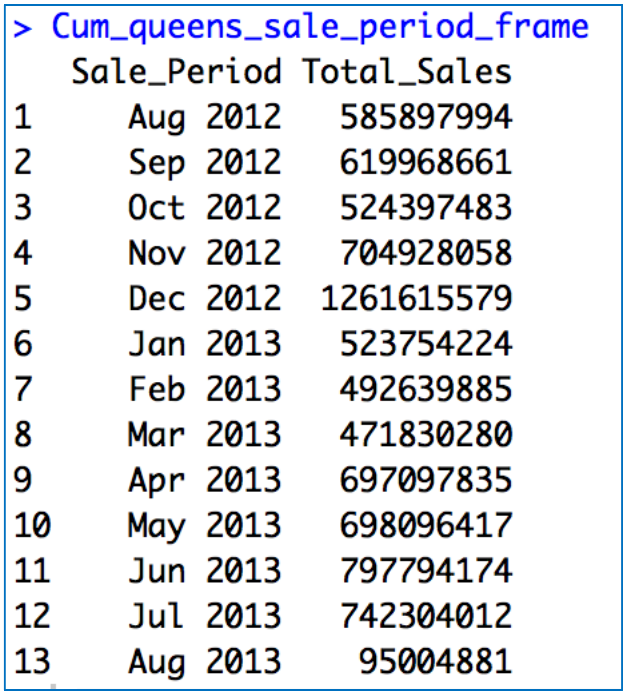
```
> Cum_manhattan_sale_period_frame
   Sale_Period Total_Sales
1     Aug 2012  3156456343
2     Sep 2012  2431564752
3     Oct 2012  3501959004
4     Nov 2012  3055128566
5     Dec 2012  9767822979
6     Jan 2013  1970663705
7     Feb 2013  1699322318
8     Mar 2013  5194577564
9     Apr 2013  2612873487
10    May 2013  3225134016
11    Jun 2013  4800550442
12    Jul 2013  2948817514
13    Aug 2013  1045684307
```

Cumulative Sales Report of Bronx



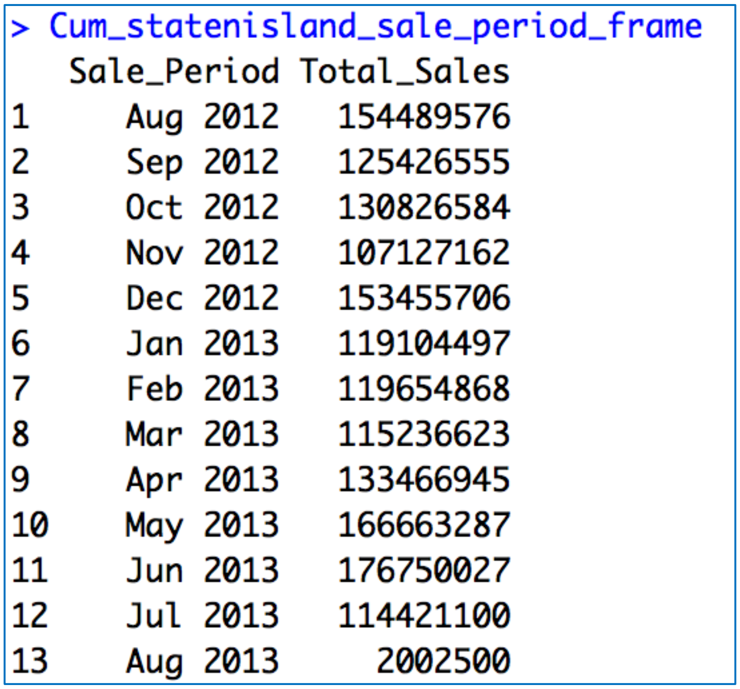
```
> Cum_bronx_sale_period_frame
   Sale_Period Total_Sales
1     Aug 2012   288923568
2     Sep 2012   155982875
3     Oct 2012   212528548
4     Nov 2012   190602548
5     Dec 2012   569294931
6     Jan 2013   102444352
7     Feb 2013   156180170
8     Mar 2013   166035985
9     Apr 2013   160981961
10    May 2013   192790621
11    Jun 2013   274493826
12    Jul 2013   266120898
13    Aug 2013    14165132
```

Cumulative Sales Report of Queens



```
> Cum_queens_sale_period_frame
   Sale_Period Total_Sales
1     Aug 2012   585897994
2     Sep 2012   619968661
3     Oct 2012   524397483
4     Nov 2012   704928058
5     Dec 2012  1261615579
6     Jan 2013   523754224
7     Feb 2013   492639885
8     Mar 2013   471830280
9     Apr 2013   697097835
10    May 2013   698096417
11    Jun 2013   797794174
12    Jul 2013   742304012
13    Aug 2013    95004881
```

Cumulative Sales Report of Staten Island

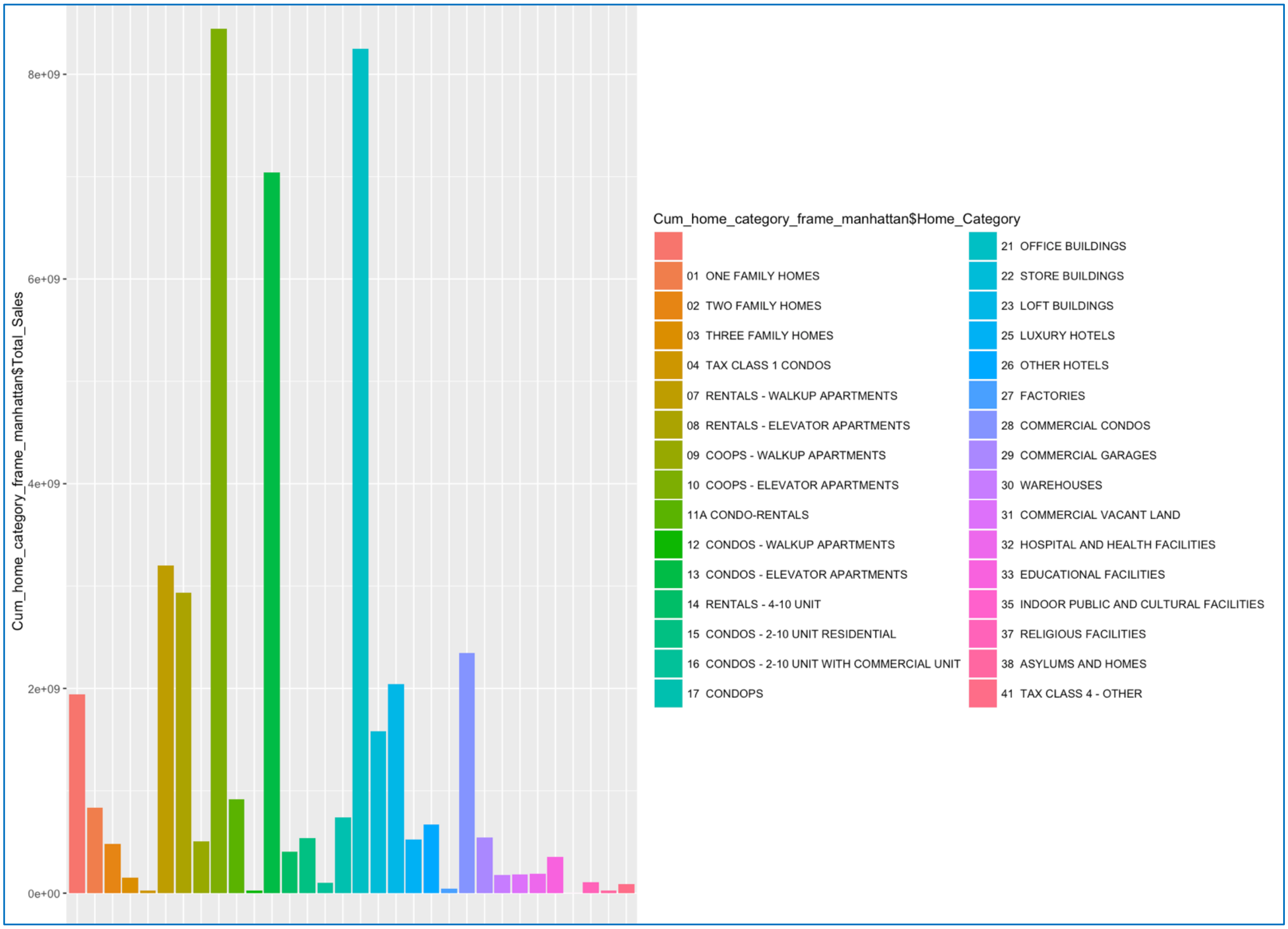


```
> Cum_statenisland_sale_period_frame
   Sale_Period Total_Sales
1     Aug 2012   154489576
2     Sep 2012   125426555
3     Oct 2012   130826584
4     Nov 2012   107127162
5     Dec 2012   153455706
6     Jan 2013   119104497
7     Feb 2013   119654868
8     Mar 2013   115236623
9     Apr 2013   133466945
10    May 2013   166663287
11    Jun 2013   176750027
12    Jul 2013   114421100
13    Aug 2013     2002500
```

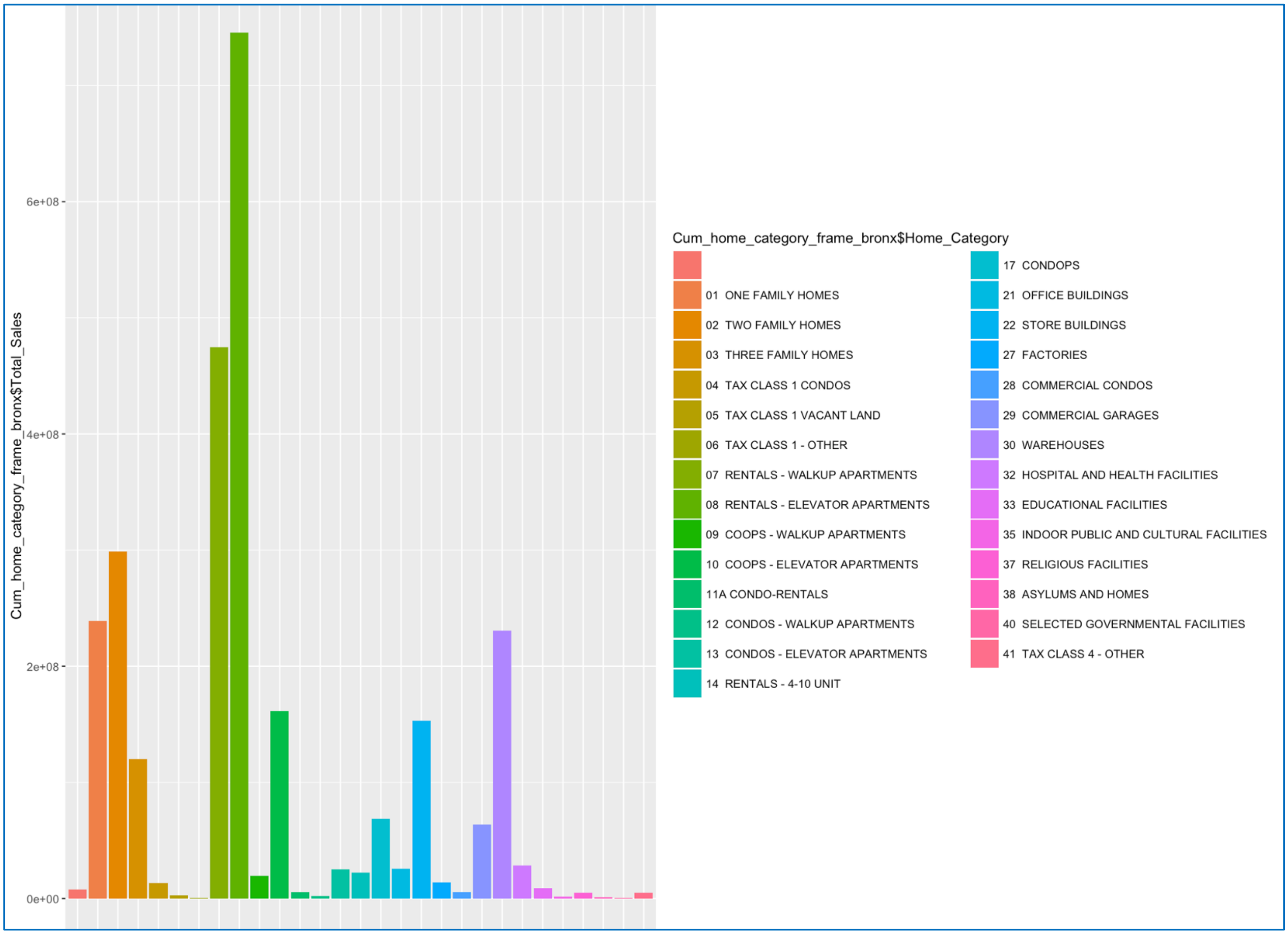
Step 3: Generating Graphs

Now we will plot graphs for the generated sales and analyze the cost of living

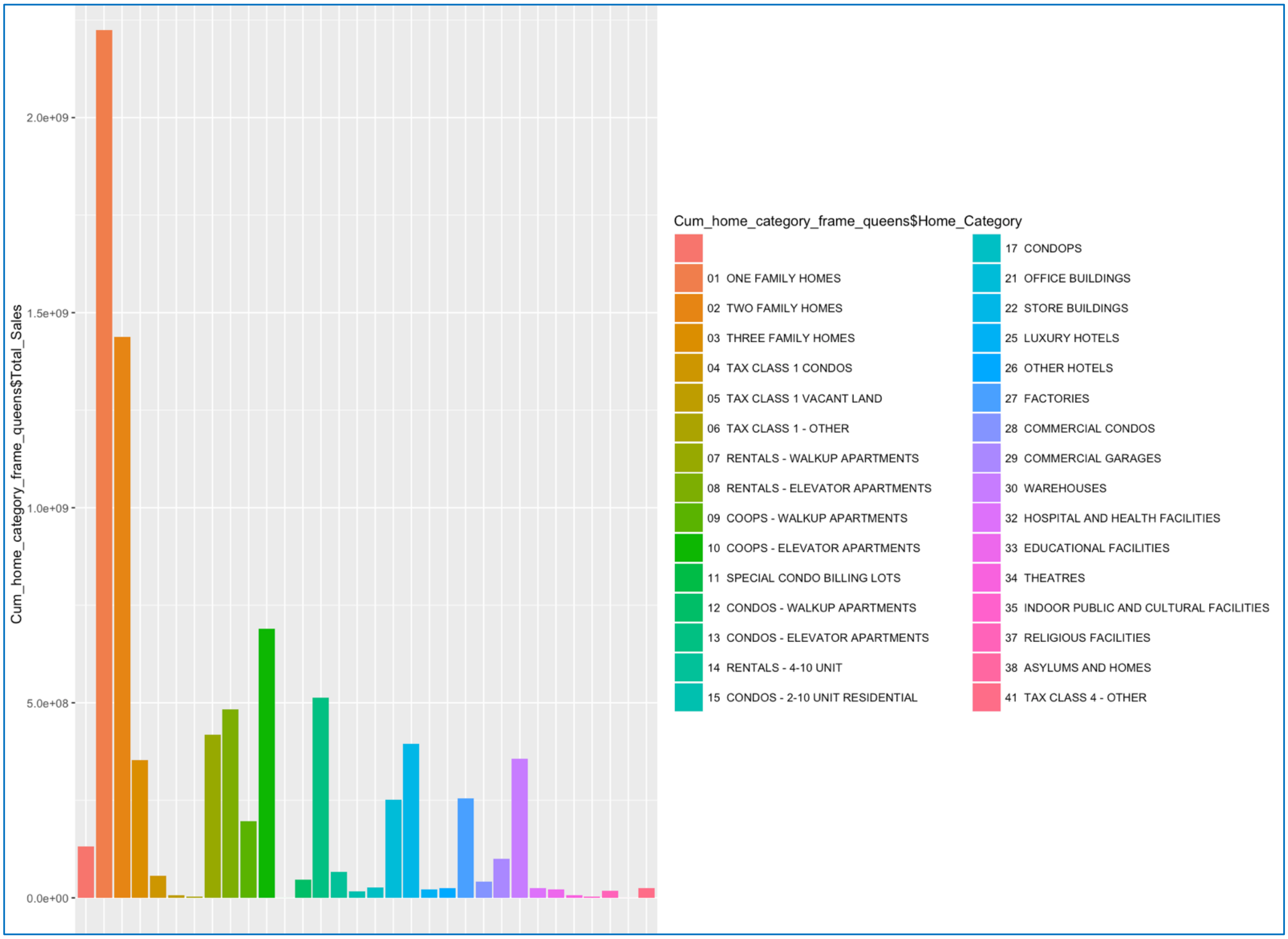
Graph on Cumulative Sales Report of Manhattan



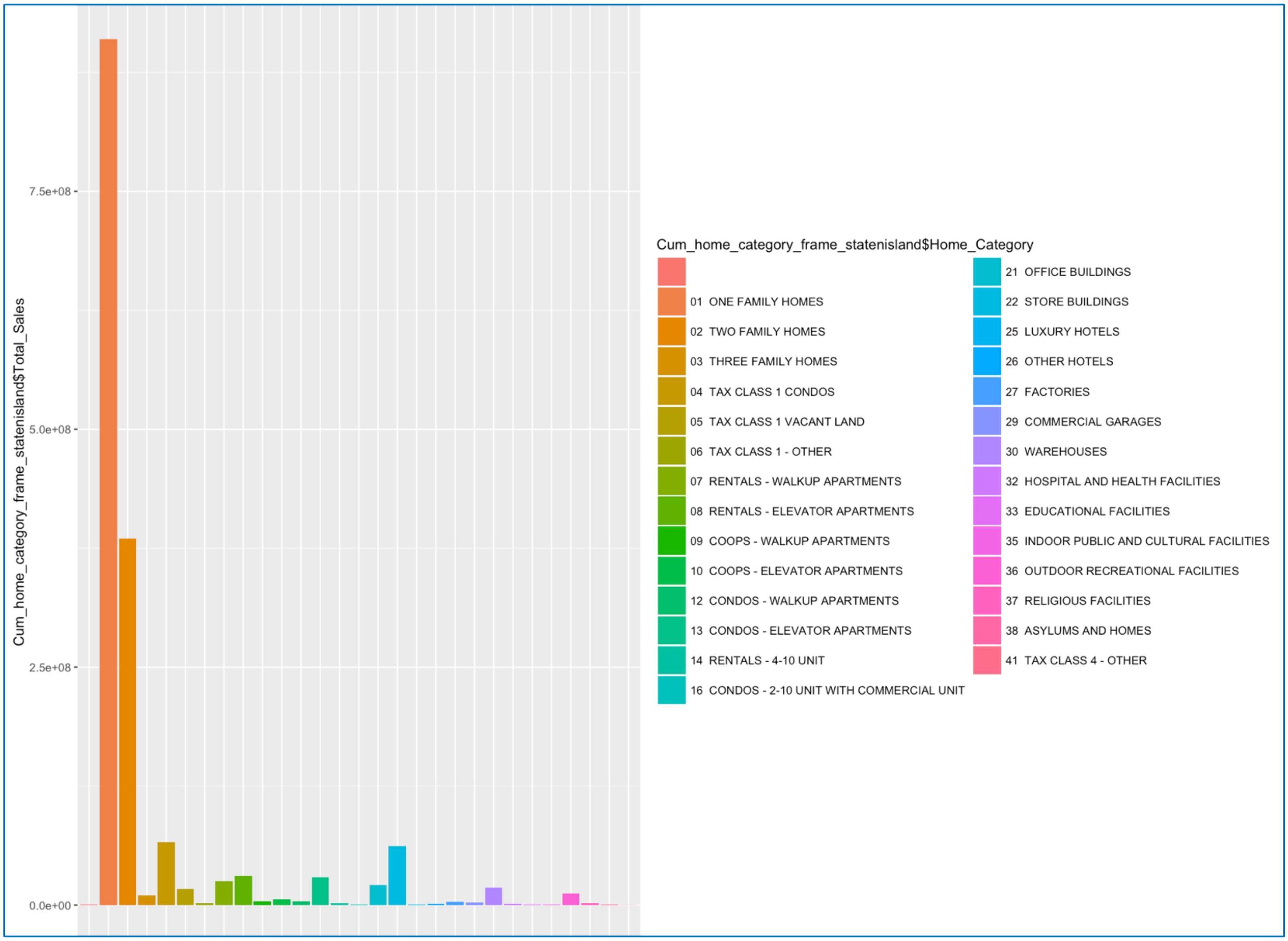
Graph on Cumulative Sales Report of Bronx



Graph on Cumulative Sales Report of Queens



Graph on Cumulative Sales Report of Staten Island


Cum_home_category_frame_statenisland$Total_Sales
7.5e+08
5.0e+08
2.5e+08
0.0e+00
Cum_home_category_frame_statenisland$Home_Category
01 ONE FAMILY HOMES
02 TWO FAMILY HOMES
03 THREE FAMILY HOMES
04 TAX CLASS 1 CONDOS
05 TAX CLASS 1 VACANT LAND
06 TAX CLASS 1 - OTHER
07 RENTALS - WALKUP APARTMENTS
08 RENTALS - ELEVATOR APARTMENTS
09 COOPS - WALKUP APARTMENTS
10 COOPS - ELEVATOR APARTMENTS
12 CONDOS - WALKUP APARTMENTS
13 CONDOS - ELEVATOR APARTMENTS
14 RENTALS - 4-10 UNIT
16 CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT
21 OFFICE BUILDINGS
22 STORE BUILDINGS
25 LUXURY HOTELS
26 OTHER HOTELS
27 FACTORIES
29 COMMERCIAL GARAGES
30 WAREHOUSES
32 HOSPITAL AND HEALTH FACILITIES
33 EDUCATIONAL FACILITIES
35 INDOOR PUBLIC AND CULTURAL FACILITIES
36 OUTDOOR RECREATIONAL FACILITIES
37 RELIGIOUS FACILITIES
38 ASYLUMS AND HOMES
41 TAX CLASS 4 - OTHER